STAT/BIOST 571: Homework 5

To be handed in on Wednesday, February 19. Where solutions require use of R, summarize your findings in a written answer, and append your annotated code, to show what you did. For each question, write up your solution on your own, using full sentences.

Problem 1: Quasilikelihood methods for the general linear model

This question examines the effect of different correlation structures, designs, and sample sizes in fitting a general linear model in a quasi-likelihood framework (i.e., assuming we know the parametric form of cluster-level covariance matrices). 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 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). 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 8, 20, or 60**
- **The design:**
  - **Design I has** $m_i = 4$, **for all clusters. In each cluster, we see** $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}\} = \{8, 10, 12, 14\}$
  - **Design II has** $m_i = 3$ **for all clusters. We see equal numbers of clusters with** $\{x_{i1}, x_{i2}, x_{i3}\} = \{8, 10, 12\}, \{8, 10, 14\}, \{8, 12, 14\},$ **or** $\{10, 12, 14\}$
- **The true covariance and the assumed covariance matriices 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.**

Present results in the table below, and write a paragraph or two summarizing your findings.

			Coverage			$\mathrm{Var}(\hat{\beta}_1)$			Relative Efficiency		
				Assumed Corr			Assumed Corr			Assumed Corr	
n	Design	True Corr	Uncor	Exch	Expon	Uncor	Exch	Expon	Uncor	Exch	Expon
8	I	Exchangeable $\rho = 0.5$	0.986	0.946	0.993	0.00327	0.00319	0.00317	0.971	0.996	1
8	I	Exchangeable $\rho = 0.9$	1	0.932	1	0.000617	0.000615	0.000643	0.996	1	0.957
8	I	Exponential $\rho = 0.5$	0.925	0.899	0.937	0.00673	0.00634	0.0062	0.921	0.978	1
8	I	Exponential $\rho = 0.9$	0.973	0.804	0.969	0.00331	0.00348	0.00323	0.975	0.929	1
8	II	Exchangeable $\rho = 0.5$	0.967	0.913	0.973	0.00543	0.00453	0.00507	0.835	1	0.894
8	II	Exchangeable $\rho = 0.9$	0.991	0.92	0.997	0.00342	0.000136	0.00912	0.266	1	0.67
8	II	Exponential $\rho = 0.5$	0.923	0.902	0.923	0.0085	0.00846	0.0084	0.987	0.992	1
8	II	Exponential $\rho = 0.9$	0.953	0.839	0.965	0.00567	0.00432	0.00418	0.737	0.967	1
20	I	Exchangeable $\rho = 0.5$	0.992	0.939	0.986	0.00129	0.00127	0.00137	0.984	1	0.926
20	I	Exchangeable $\rho = 0.9$	1	0.954	1	0.000242	0.000239	0.00027	0.988	1	0.887
20	I	Exponential $\rho = 0.5$	0.935	0.913	0.935	0.00276	0.00267	0.00271	0.968	1	0.989
20	I	Exponential $\rho = 0.9$	0.981	0.832	0.992	0.00146	0.00132	0.00125	0.859	0.947	1
20	II	Exchangeable $\rho = 0.5$	0.978	0.938	0.985	0.00217	0.00185	0.00189	0.853	1	0.98
20	II	Exchangeable $\rho = 0.9$	0.997	0.937	1	0.000385	0.00122	0.000525	0.316	1	0.734
20	II	Exponential $\rho = 0.5$	0.931	0.902	0.933	0.00348	0.0037	0.00346	0.996	0.935	1
20	II	Exponential $\rho = 0.9$	0.967	0.882	0.983	0.00247	0.0017	0.00162	0.658	0.955	1
60	I	Exchangeable $\rho = 0.5$	0.994	0.95	0.988	0.000411	0.000395	0.000462	0.959	1	0.854
60	I	Exchangeable $\rho = 0.9$	1	0.944	1	8.75e-05	7.85e-05	8.82e-05	0.897	1	0.891
60	I	Exponential $\rho = 0.5$	0.942	0.906	0.946	0.000888	0.000967	0.00088	0.998	0.918	1
60	I	Exponential $\rho = 0.9$	0.99	0.854	0.995	0.000453	0.00044	0.000421	0.929	0.956	1
60	II	Exchangeable $\rho = 0.5$	0.98	0.941	0.988	0.00072	0.000605	0.00067	0.841	1	0.904
60	II	Exchangeable $\rho = 0.9$	0.997	0.946	1	0.000459	0.000122	0.000173	0.266	1	0.706
60	II	Exponential $\rho = 0.5$	0.947	0.924	0.95	0.00118	0.00115	0.00108	0.919	0.938	1
60	II	Exponential $\rho = 0.9$	0.974	0.865	0.993	0.000772	0.000582	0.000576	0.747	0.99	1

Table of true coverage (percentage), at the nominal 95% level, variance of the estimated slopes, and efficiency of the estimates of the slope. Efficiencies are given relative to the most efficient estimator, in each row

We simulated 1000 times for each entry of the table, where row differes mainly in the true correlation structure that generates the data; column differes in the assumption on structure of working correlation matrix. Especially, when we assume working independence, we should get same estimate as OLS.

Comparing the coverages by column, we can see that the coverage is closest to 95% when we correctly specify the correlation structure. Moreover, under independence assumption, the coverage tend to be larger than 95%; under exchangeable working covariance assumption, when the true structure is exponential, the coverage tend to be smaller than 95%; under exponential working covariance assumption, when the true structure is exchangeable, the coverage tend to be larger

than 95% (which we would be more concerned about).

Comparing the coverages by row, we can see that when \rho is 0.9, the trend discussed above will be more extreme. We also see that the design (balanced versus unbalanced) doesn't change the result too much.

Comparing the true variance of βˆ (or the relative efficiency) by column, we can see that, the true variance 1 is most effiecient when we correctly specified the correlation structure. Also, the uncorrelated structure tends to yield a relatively more variable estimate of $\beta1$, as we omitted the information from taking correlation within cluster into account.

Comparing the true variance of βˆ (or the relative efficiency) by row, we again see a more extreme differemce 1 when there is much more information to explore from within cluster correlation, i.e. when $\rho = 0.9$. Moreover, the design (balanced versus unbalanced) doesn't change the result too much.

Problem 2: Semiparametric methods for the general linear model

Extend your simulations from Problem 1 by adding semi-parametric sandwich standard errors that account for clustering. As before, write your own code "by hand" using matrix algebra. The focus of this exercise is to compare two types of standard error estimates and the true standard error of $\hat{\beta}_1$.

Present results in the table below, and write a paragraph or two summarizing your findings.

			$SD(\hat{\beta}_1)$			$E(\widehat{SE}_{1,QL})$			$E(\widehat{SE}_{1,sand})$		
			Assumed Corr			**Assumed Corr**			**Assumed Corr**		
n	**Design**	**True Corr**	**Uncor**	**Exch**	**Expon**	**Uncor**	**Exch**	**Expon**	**Uncor**	**Exch**	**Expon**
8	**I**	**Exchangeable** $\rho = 0.5$	**0.0572**	**0.0564**	**0.0563**	**0.0747**	**0.0544**	**0.0743**	**0.0506**	**0.051**	**0.0516**
8	**I**	**Exchangeable** $\rho = 0.9$	**0.0248**	**0.0248**	**0.0253**	**0.0719**	**0.0245**	**0.0644**	**0.0223**	**0.0228**	**0.0232**
8	**I**	**Exponential** $\rho = 0.5$	**0.082**	**0.0796**	**0.0787**	**0.075**	**0.0701**	**0.0773**	**0.0748**	**0.0732**	**0.074**
8	**I**	**Exponential** $\rho = 0.9$	**0.0576**	**0.059**	**0.0568**	**0.0736**	**0.0404**	**0.0707**	**0.0525**	**0.0527**	**0.0513**
8	**II**	**Exchangeable** $\rho = 0.5$	**0.0737**	**0.0673**	**0.0712**	**0.0844**	**0.0639**	**0.0838**	**0.0663**	**0.0597**	**0.0617**
8	**II**	**Exchangeable** $\rho = 0.9$	**0.0585**	**0.0302**	**0.0369**	**0.083**	**0.0287**	**0.0721**	**0.0501**	**0.0266**	**0.0319**
8	**II**	**Exponential** $\rho = 0.5$	**0.0922**	**0.092**	**0.0916**	**0.0854**	**0.08**	**0.0871**	**0.0833**	**0.0809**	**0.0814**
8	**II**	**Exponential** $\rho = 0.9$	**0.0753**	**0.0657**	**0.0646**	**0.0831**	**0.0481**	**0.0789**	**0.0664**	**0.0569**	**0.0578**
20	**I**	**Exchangeable** $\rho = 0.5$	**0.036**	**0.0357**	**0.0371**	**0.0486**	**0.0349**	**0.0492**	**0.0342**	**0.0339**	**0.0346**
20	**I**	**Exchangeable** $\rho = 0.9$	**0.0156**	**0.0155**	**0.0164**	**0.0485**	**0.0156**	**0.0445**	**0.0154**	**0.0152**	**0.0158**
20	**I**	**Exponential** $\rho = 0.5$	**0.0526**	**0.0517**	**0.052**	**0.0491**	**0.0456**	**0.0505**	**0.0501**	**0.0501**	**0.0497**
20	**I**	**Exponential** $\rho = 0.9$	**0.0382**	**0.0363**	**0.0354**	**0.0486**	**0.0262**	**0.0479**	**0.0351**	**0.0348**	**0.0347**
20	**II**	**Exchangeable** $\rho = 0.5$	**0.0466**	**0.043**	**0.0434**	**0.0558**	**0.0419**	**0.0558**	**0.0449**	**0.0409**	**0.0424**
20	**II**	**Exchangeable** $\rho = 0.9$	**0.0349**	**0.0196**	**0.0229**	**0.0555**	**0.0188**	**0.05**	**0.0349**	**0.0184**	**0.022**
20	**II**	**Exponential** $\rho = 0.5$	**0.059**	**0.0609**	**0.0588**	**0.0567**	**0.0527**	**0.0574**	**0.0567**	**0.0556**	**0.0561**
20	**II**	**Exponential** $\rho = 0.9$	**0.0497**	**0.0412**	**0.0403**	**0.0558**	**0.0318**	**0.0541**	**0.0456**	**0.0389**	**0.0396**
60	**I**	**Exchangeable** $\rho = 0.5$	**0.0203**	**0.0199**	**0.0215**	**0.0287**	**0.0203**	**0.029**	**0.0202**	**0.02**	**0.0206**
60	**I**	**Exchangeable** $\rho = 0.9$	**0.00936**	**0.00886**	**0.00939**	**0.0285**	**0.00909**	**0.0266**	**0.00903**	**0.00903**	**0.00932**
60	**I**	**Exponential** $\rho = 0.5$	**0.0298**	**0.0311**	**0.0298**	**0.0287**	**0.0266**	**0.0296**	**0.0295**	**0.0296**	**0.0294**
60	**I**	**Exponential** $\rho = 0.9$	**0.0213**	**0.021**	**0.0205**	**0.0285**	**0.0153**	**0.0283**	**0.0208**	**0.0207**	**0.0206**
60	**II**	**Exchangeable** $\rho = 0.5$	**0.0268**	**0.0246**	**0.0259**	**0.033**	**0.0245**	**0.0329**	**0.0268**	**0.0243**	**0.025**
60	**II**	**Exchangeable** $\rho = 0.9$	**0.022**	**0.0112**	**0.0131**	**0.0329**	**0.0111**	**0.0301**	**0.0207**	**0.011**	**0.0131**
60	**II**	**Exponential** $\rho = 0.5$	**0.0343**	**0.034**	**0.0329**	**0.033**	**0.031**	**0.0336**	**0.0335**	**0.0333**	**0.0334**
60	**II**	**Exponential** $\rho = 0.9$	**0.0278**	**0.0241**	**0.024**	**0.0329**	**0.0186**	**0.0321**	**0.0272**	**0.0229**	**0.0237**

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

Same as problem 1, we simulated 1000 times for each entry of the table, where row differes mainly in the true correlation structure that generates the data; column differes in the assumption on structure of working correlation matrix. The three major blocks of columns are: the true standard deviation, quasi-likelihood standard error and sandwich standard error. Especially, when we assume working independence, we should get same beta estimate as OLS.

From the table we can see that, comparing column-wise

(1) when we correctly specify the working correlation matrix, the estimated standard error from quasi-likelihood standard error is close to the truth; when we are wrong about the correlation structure, the quasi-likelihood standard error is different from the truth.

(2) sandwich standard error is robust to misspecification of the working correlation matrix, and the third column is similar to the first column. Especially, as n grows, the difference gets smaller;

when n is relatively small (n=8), sandwich standard error underestimates the truth.

(3) when we use working independence matrix, the standard error is larger than those from the other two correlation structures.

Comparing row-wise

(1) as n grows, the standard error tends to get smaller.

(2) when correlation is large (0.9), the standard error is smaller.

(3) standard errors from design I is generally smaller thatn from design II.