Biostatistics 140.655, 2017-18
QUIZ 2

Quiz Guidelines:

Please read the following quiz guidelines carefully:

- For this quiz, you are to work ALONE. You may use your course notes and lab materials to help answer the questions.
- Submit your answers to Courseplus on Friday February 9th by 5pm.
- DO NOT discuss this quiz or your solution to this quiz with other students from the course on Wednesday Feb 7th, Thursday Feb 8th or Friday Feb 9th. The solution to the quiz will be available Saturday, Feb 10th.
- By submitting your answers to Courseplus, you are acknowledging that you have read the guidelines carefully and will adhere to these guidelines.

Scientific Background:

We will be exploring the variance and within subject correlation patterns within the data you are analyzing for Homework 1. The dataset contains a random sample of 300 girls living in Topeka, Kansas from the Six Cities Study of Air Pollution and Health.

Recall that the majority of the children in the study were enrolled in the first or second grade (between the ages of six and eight) but many children were enrolled at ages over eight years. After the first assessment, measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian. In addition, age and height were measured.

Notation:

Let Y_{ij} be the Log(FEV1) measured at annual visit j for girl i, where $i = 1, \ldots, 300$ and $j = 1, \ldots, n_i$ where the $\max(n_i) = 12$. The two primary exposure variables are age_{ij} and $height_{ij}$.

We will assume $\boldsymbol{Y_i}$ is the vector of responses for girl i and that $\boldsymbol{Y_i} \sim \text{MVN}(\mu_i, V_i)$.

Objective:

For this quiz, we will focus on evaluating the structure of $\boldsymbol{V_i}$.

To evaluate the structure of $\boldsymbol{V_i}$, we will create residuals where we remove the effects of age and height from the Log(FEV1).

Specifically, we do the following:

```
xtset id visit
gen age2 = age^2
gen height2 = height^2
quietly regress logfev1 age age2 height height2
predict logfev_res, residual
```

1. To understand the structure of V_i, it is important to explore both the within subject correlation structure and the variation in Log(FEV1) at each follow-up time. To explore the variation in Log(FEV1) at each visit, you could
    a. Make a scatterplot of the residuals of Log(FEV1) vs. visit. The variance is roughly constant if the mean residual at each visit is zero.
    b. Make a scatterplot of the residuals of Log(FEV1) vs. visit. The variance is roughly constant if the range of the majority of the residuals is roughly equal at each visit.

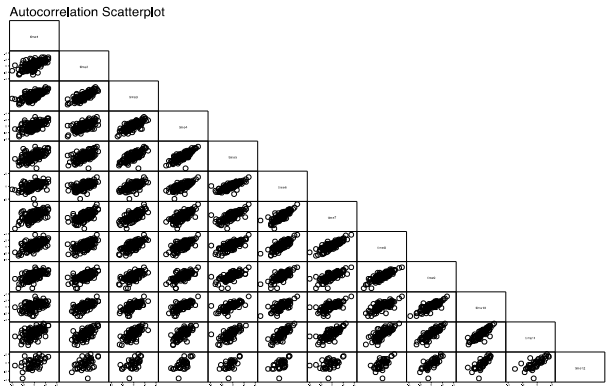
Next, you explore the within subject correlation structure by estimating both the empirical correlation matrix and autocorrelation function.

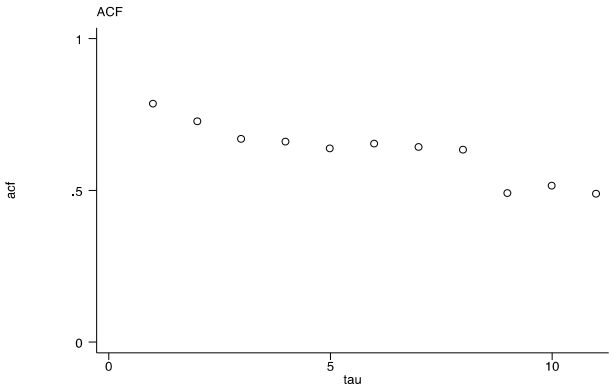
```
autocor logfev_res visit id
```

	time1	time2	time3	time4	time5	time6	time7
time1	1.0000						
time2	0.6646	1.0000					
time3	0.6878	0.8078	1.0000				
time4	0.6154	0.6620	0.7809	1.0000			
time5	0.5637	0.5977	0.7310	0.8145	1.0000		
time6	0.5120	0.5848	0.6590	0.7336	0.7919	1.0000	
time7	0.5968	0.6546	0.7418	0.7291	0.7617	0.8311	1.0000
time8	0.6242	0.6264	0.7141	0.7227	0.7257	0.7978	0.8415
time9	0.6096	0.6499	0.6694	0.6886	0.7197	0.7565	0.7643
time10	0.4339	0.6632	0.6705	0.7316	0.7276	0.7400	0.7007
time11	0.5639	0.6816	0.7036	0.7313	0.7027	0.7990	0.7548
time12	0.4856	0.4398	0.5175	0.5918	0.6133	0.5433	0.4118

	time8	time9	time10	time11	time12
time8	1.0000				
time9	0.8474	1.0000			
time10	0.7390	0.8270	1.0000		
time11	0.7877	0.8546	0.8566	1.0000	
time12	0.6177	0.7559	0.7650	0.7226	1.0000

	acf
1.	.7844789
2.	.7263785
3.	.66692
4.	.6582604
5.	.6372369
6.	.6522721
7.	.6410289
8.	.6309119
9.	.4885103
10.	.5123967
11.	.4855894





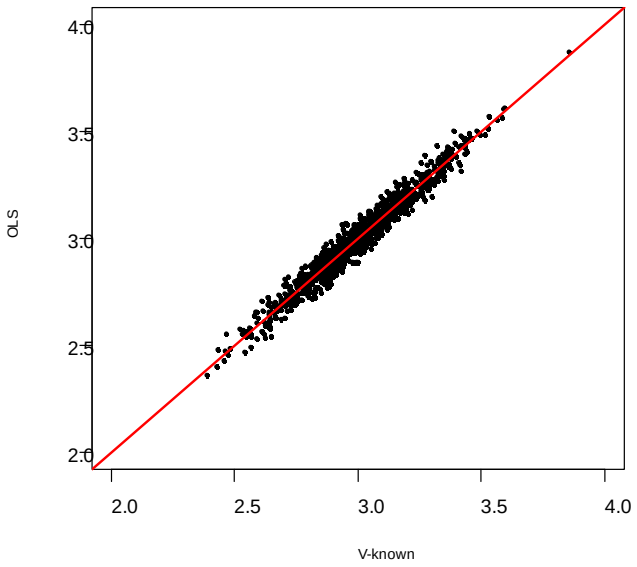
2. Based on the estimated within-subject correlation matrix, the estimated correlation between Log(FEV1) values measured at visit 3 and 6 is
    a. 0.6878
    b. 0.6669
    c. 0.6590
    d. 0.5120

3. Based on the estimated autocorrelation function, the estimated correlation between Log(FEV1) values measured at visit 3 and 6 is
    a. 0.6878
    b. 0.6669
    c. 0.6590
    d. 0.5120

4. Based on the estimated autocorrelation function, a reasonable parametric model for the within subject correlation would be (there may be two answers below that would be "reasonable", please select only one answer)
    a. Unstructured
    b. Exchangeable
    c. Toeplitz
    d. Autoregressive 1

5. Suppose you fit a model for Log(FEV1) where the mean Log(FEV1) is explained by age (linear) and height (linear) and the model for V_i assumes constant variance at each visit and an autoregressive(1) correlation structure. Suppose the estimate of the correlation parameter was 0.79. Using this parametric correlation model, the estimated correlation between Log(FEV1) values measured 5 years apart
    a. is similar to that estimated by the autocorrelation function
    b. overestimates the correlation compared to that estimated by the autocorrelation function
    c. underestimates the correlation compared to that estimated by the autocorrelation function
    d. cannot be compared to that estimated by the autocorrelation function.

In lab 2 we evaluated the results of a simulation study. The goal was to better understand the behavior of regression coefficients estimated using weighted least squares where we may incorrectly define V_i, the variance matrix for subject *i*. The figure below compares the estimated monthly improvement in SF-36 mental health scores based on samples of size 100 patients across 10,000 hypothetical studies. The x-axis displays the estimated monthly improvement in SF-36 mental health scores when we assumed the correct model for V_i (x-axis) and the y-axis displays the same estimates where we assumed the observations within a subject over time were uncorrelated (ordinary least squares model).



6. Based on the estimates of bias we reviewed in the lab we found that when we average over possible samples of size 100, the OLS estimator provides an unbiased estimate for the true monthly improvement in SF-36 mental health scores. When we compare the estimated monthly improvement in SF-36 mental health scores generated by OLS vs. the true WLS model,
    a. We note that in any individual sample, the OLS estimate may be larger or smaller than the corresponding estimate from the true WLS model.
    b. We note that in any individual sample, the OLS estimate will be larger than the corresponding estimate from the true WLS model.
    c. We note that in any individual sample, the OLS estimate will be smaller than the corresponding estimate from the true WLS model.

In lab 2, we generated balanced designs (the same number of observations per subject) where the goal was to estimate the linear slope for time and the within subject correlation model was autoregressive 1 with $\boldsymbol{\rho = 0.9}$. You computed the relative efficiency comparing the variance of the estimated monthly improvements in SF-36 mental health scores based on the OLS vs. the true WLS model. When $\boldsymbol{\rho = 0.9}$, the relative efficiency is roughly 1.05; i.e. the variance of the estimated monthly improvement is 5 percent greater if you use the OLS model relative to the true WLS model.

7. The relative efficiency will depend on the strength of the correlation. Suppose the true correlation model was autoregressive 1 with $\boldsymbol{\rho = 0.3}$, which of the following is your best guess at the relative efficiency when $\boldsymbol{\rho = 0.3}$.
    a. 1.00
    b. 1.02
    c. 1.05
    d. 1.10

FINAL Question 8 is on the next page!

8. Some would argue that a 5 percent difference in the variance of an estimate is negligible, so why not always use the OLS model in a balanced design. The answer to this question is that in balanced designs, we do not always see such small losses in efficiency. Consider the following example extracted from Chapter 4 of the Diggle, Heagerty, Liang and Zeger text.

   Suppose you will implement a two-treatment crossover design in which $n = 3$ measurements are taken at months 1, 2, and 3 on each of $m = 8$ subjects. The sequences of treatments given to the subjects are AAA, AAB, ABA, ABB, BAA, BAB, BBA and BBB. The true model for the data is given by

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 I(treatmentB)_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)\ and\ Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho^{|j-k|}}$$

   The relative efficiencies for estimation of $\boldsymbol{\beta_1}$, i.e. ratio of the variance of the estimate for $\boldsymbol{\beta_1}$ comparing the OLS to the true WLS model, are given below. With $\boldsymbol{\rho = 0.9}$, the variance of the estimate $\boldsymbol{\beta_1}$ based on the OLS model is more than 500% larger than the variance of the estimate of $\boldsymbol{\beta_1}$ based on the WLS model.

$\boldsymbol{\rho}$	0.1	0.3	0.5	0.7	0.9
Relative efficiency	1.01	1.13	1.44	2.28	6.67

   We note that the key difference between our simulation study in lab 2 and this example is:

   a. The data are balanced in the design we considered in lab 2; whereas in this example the data are not balanced.
   b. The model for the within subject correlation structure differ which may be causing the differences in the relative efficiencies.
   c. The interpretation of $\boldsymbol{\beta_1}$ are different. In lab 2, $\boldsymbol{\beta_1}$ is the monthly improvement in SF-36 mental health scores (i.e. a linear slope for time). The goal from the lab is to estimate the within subject monthly change in SF-36 mental health scores over time. In the example from the textbook, $\boldsymbol{\beta_1}$ is the difference in the mean response comparing treatment B to treatment A. The goal is to estimate the within subject effect of receiving treatment B relative to treatment A. Therefore, the differences in the efficiency can be attributable to differences in the mean model specification or the parameters of interest.