

**Biostatistics 140.655, 2017-18**  
**Lab 2**

**Topics:**

- Autocorrelation function
- Parametric models for the within subject correlation
- Weighted least squares
- Comparison of the bias and variance in estimating a linear slope using various modeling assumptions about the within subject covariance structure.

**Learning Objectives:**

Students who successfully complete this lab will be able to:

- Fit the saturated model for data with a single factor (time) and obtain residuals
- Estimate the sample autocorrelation function
- Deduce a parametric model for the within subject correlation based on the sample autocorrelation function
- Fit a weighted least squares regression model
- Describe statistical properties of the weight least squares solution

**Associated Quiz:**

- While we will review and discuss parts of this exercise, there is a short quiz (Quiz 2) on Courseplus which will assess your basic knowledge of the course materials thus far with focus on ideas from this lab session.
- Quiz 2 will be available on Wednesday Feb 7<sup>th</sup>, 5pm.
- Submit your quiz solution by 5pm Friday Feb 9<sup>th</sup>.

**Scientific Background (same as before):**

Assume you are a researcher interested in mental health symptoms among critically ill ICU survivors. You administered the Short Form (36) Health Survey (SF-36) to 100 patients that consented to participate in your study. The SF-36 will be administered at hospital discharge (time 0) and then monthly for 4 months. You are specifically interested in the mental health score of the SF-36.

*A priori* you believe that the mental health symptoms of the ICU survivors will improve over the course of the follow-up and you state that you will estimate the improvement in mental health symptoms comparing 1 to 4 months post hospital discharge to hospital discharge (time 0 or baseline).

NOTE: We are going to assume we have no deaths in our study patients or drop-out/missing data. We will address these issues later in the course.

**Lab Exercise:**

1. You will explore the autocorrelation function within three hypothetical studies.

Each of the three hypothetical studies was generated assuming the following:

$$\begin{pmatrix} Y_{i0} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} \sim MVN \left( \begin{pmatrix} 35 \\ 38 \\ 41 \\ 44 \\ 47 \end{pmatrix}, \begin{pmatrix} 100 & ? & ? & ? & ? \\ ? & 100 & ? & ? & ? \\ ? & ? & 100 & ? & ? \\ ? & ? & ? & 100 & ? \\ ? & ? & ? & ? & 100 \end{pmatrix} \right)$$

Your goal is to fill in the “?” within the variance specification for this k-variate normal distribution (k = 5). To fill in the “?” you will be identifying the parametric model that defines the within subject correlation.

Go to the Courseplus site and download the lab2.zip file and find data from the three hypothetical studies: “autocor1”, “autocor2” and “autocor3”. The data are saved in both Stata format (.dta extension) and as comma delimited files (.csv extension) for those using R.

Fill in the following table:

	Hypothetical Study 1	Hypothetical Study 2	Hypothetical Study 3
Lag	Sample autocorrelation function	Sample autocorrelation function	Sample autocorrelation function
1			
2			
3			
4			
Parametric model:			

**STATA:**

```
regress y i.time
predict resid, resid
autocor resid time id
```

**R:**

```
fit <- gls(y~as.factor(time),data)
ACF(fit,form=~1|id)
```

2. For each of the three hypothetical studies, estimate the monthly improvement in SF-36 mental health scores using both ordinary least squares and weighted least squares with an unstructured variance model; treat time as a linear variable. Fill in the table below:

Data	Monthly improvement in SF-36			
	Coefficient		SE	
	OLS	WLS	OLS	WLS
autocor1				
autocor2				
autocor3				

What is the true monthly improvement in SF-36 mental health scores?

For a given dataset, do the estimated monthly improvement in SF-36 mental health scores differ across the two statistical methods? If so, why?

For a given dataset, do the standard errors for the estimated monthly improvement in SF-36 mental health scores differ across the two statistical methods? If so, why?

### **STATA:**

```
regress y time
mixed y time || id: , nocons residuals(un, t(time))
```

### **R:**

```
fit.ols <- lm(y~time,data)
fit.wls <- gls(y~time,correlation=corSymm(form = ~ 1 | id),
data=data,method="ML",weights=varIdent(form=~1 | as.factor(time)))
```

3. Above, you compared the estimated monthly improvement in SF-36 mental health scores generated from the OLS and WLS procedures from a single study of 100 patients. Here, we will explore the repeated sampling behavior of the estimated monthly improvement in SF-36 mental health scores assuming various models for the within subject variance across increasing sample sizes. You will identify important patterns in the behavior of the estimates based on the specified model for the variance. NOTE: We are exploring the properties of WLS within the context of no missing data; we will consider this in more detail later in the course.

**SIMULATION STUDY:** Please DO NOT run the simulation study on your laptop; I ran the simulation in R on the Department of Biostatistics computing cluster using 25 parallel processing cores and it took a couple of hours to complete.

I generated 10,000 simulated studies with  $m = 10, 25, 100, 500$  and 1000 patients. The patients were sampled from a population of patients whose data follows the  $k$ -variate normal distribution ( $k = 5$ ) below:

$$\begin{pmatrix} Y_{i0} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} \sim MVN \left( \begin{pmatrix} 35 \\ 38 \\ 41 \\ 44 \\ 47 \end{pmatrix}, V_i = \begin{pmatrix} 100 & 100 \times \rho & 100 \times \rho^2 & 100 \times \rho^3 & 100 \times \rho^4 \\ 100 \times \rho & 100 & 100 \times \rho & 100 \times \rho^2 & 100 \times \rho^3 \\ 100 \times \rho^2 & 100 \times \rho & 100 & 100 \times \rho & 100 \times \rho^2 \\ 100 \times \rho^3 & 100 \times \rho^2 & 100 \times \rho & 100 & 100 \times \rho \\ 100 \times \rho^4 & 100 \times \rho^3 & 100 \times \rho^2 & 100 \times \rho & 100 \end{pmatrix} \right)$$

Therefore, the patients are sampled from a population where the monthly improvement in the SF-36 mental health scores is 3 units, the variance of the SF-36 mental health scores at any time is 100 and the correlation between any two SF-36 mental health scores is given by  $\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|j-k|}$  and  $\rho = 0.9$ ; the AR1 model.

In each of the 10,000 simulated studies, I estimated the monthly improvement in SF-36 mental health scores using the correct model for the mean (linear function of month) and the following models for the within subject variance:

- i. WLS – V known: I provided the correct information for  $V_i$ ; i.e.  $\text{Var}(Y_{ij}) = 100$  and  $\text{Corr}(Y_{ij}, Y_{ik}) = 0.9^{|j-k|}$
- ii. WLS – V estimated: I assumed the correct model for  $V_i$  but I estimated the required parameters within each of the simulated studies
- iii. WLS – V unstructured: I did not assume a model for  $V_i$  so estimated 5 variance and 10 correlation parameters within each of the simulated studies.
- iv. OLS: I assumed the SF-36 mental health scores from the same subject were uncorrelated and that the variance of the SF-36 mental health scores was the same at all the measurement times and estimated the variance within each of the simulated studies.

The table below displays the bias and variance of the 10,000 estimated monthly improvements in SF-36 mental health scores based on different sample sizes and the models i. through iv. The bias is defined as the average of the 10,000 estimated monthly improvements in SF-36 mental health scores over the 10,000 simulated studies minus 3 (the true monthly improvement).

Sample size (m)	Bias				Variance			
	WLS – V known	WLS – V estimated	WLS - unstructured	OLS	WLS – V known	WLS – V estimated	WLS - unstructured	OLS
<b>10</b>	-0.005	-0.005	-0.007	-0.003	0.422	0.422	0.678	0.442
<b>25</b>	-0.004	-0.004	-0.003	-0.003	0.172	0.172	0.200	0.181
<b>100</b>	-0.002	-0.002	-0.002	-0.002	0.0430	0.0430	0.0443	0.0451
<b>500</b>	-0.0005	-0.0005	-0.0004	-0.0003	0.00883	0.00883	0.00888	0.00925
<b>1000</b>	-0.0006	-0.0007	-0.0007	-0.0005	0.00439	0.00439	0.00440	0.00458

Based on the results in the table answer the following questions:

- For a fixed sample size, how does the bias compare across the models specified for the within subject variance?
- Regardless of the model selected for the within subject variance, how does the bias change as the sample size goes from small ( $m = 10$ ) to large ( $m = 1000$ )?
- Compare the variance across WLS – V known to WLS – V estimated for fixed sample sizes.
- For small sample sizes ( $m = 10$  and  $m = 25$ ), compare the variance for WLS – V known and WLS – V estimated to WLS – unstructured? If there are differences, what do you think drives the differences?
- For larger sample sizes ( $m = 100$  to  $m = 1000$ ), compare the variance for WLS – V known and WLS – V estimated to WLS – unstructured? If there are differences, what do you think drives the differences?
- For each of the sample sizes considered, compute the relative efficiency of estimating the monthly improvement assuming the correct variance model to assuming independence.

Some comments on the properties of WLS and frequently asked questions:

- I. Regardless of how we specify the model for  $V$ , the WLS procedure produces an unbiased estimate of the monthly improvement in SF-36 mental health scores.
- II. However, specifying the wrong model for  $V$  can have an impact on your inference! EXAMPLE: the OLS estimator produces an estimate of the variance that is 5% too large!
- III. Why not always use the unstructured approach? ANSWER: in small samples, you found that the variance for the WLS- unstructured was inflated and what defines “small samples” will change depending on the complexity of the mean model and  $n$  (the number of observations within a subject). In addition, sometimes there are unexpected dependencies in the empirical estimate of  $V$  using the unstructured approach so the model doesn’t fit (i.e. we are unable to invert the estimated  $V$ ).
- IV. Why not always use OLS? You could if your goal is only estimation (i.e. interest is in prediction) not inference (hypothesis testing, confidence intervals). If you are interested in inference, then you can get the wrong answer!
- V. Our simulation study focused on data generated from an underlying multivariate normal distribution. After we estimate the monthly improvement in SF-36 mental health scores, we want to estimate a confidence interval. Our confidence interval methods rely on the assumption that the slope estimate is normally distributed. Even if the data are not normally distributed, the normality of the slope estimates hold in large samples, due to the central limit theorem.