Biostatistics 140.655, 2017-18
Lab 1 Solution

Topics:

- Generate multivariate normal data
- Comparison of the width and confidence level of confidence intervals constructed with and without accounting for the correlation in the data
- Exploratory analysis of the mean response profile within a longitudinal dataset
- Exploratory analysis of the within subject correlation within a longitudinal dataset

Learning Objectives:

Students who successfully complete this lab will be able to:

- Compute the sample mean and variance of longitudinal data and the sample correlation between two observations from the same subject measured at different times
- Create figures describing the mean response profile within a longitudinal dataset
- Define a "saturated model" for the mean
- Estimate the autocorrelation function and describe each element of the autocorrelation function

Associated Quiz:

- Quiz 1 will be available on Courseplus at 5pm today (Feb 1st).
- You are to work alone on the quiz but may consult your course notes and lab materials.
- Submit your solution to the quiz via Courseplus by end of day Feb 3rd.

Scientific Background:

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.

To be completed prior to lab on Wednesday:

STATA Users: To estimate the autocorrelation function in Stata, you need to install a new program called "autocor".

- Open a Stata session.
- Find out where Stata is storing ".ado" files on your computer by typing "adopath"

  Here is what I get on my PC:

```
. adopath
  [1]  (BASE)      "C:\Program Files (x86)\Stata13\ado\base/"
  [2]  (SITE)      "C:\Program Files (x86)\Stata13\ado\site/"
  [3]              "."
  [4]  (PERSONAL)  "c:\ado\personal/"
  [5]  (PLUS)      "c:\ado\plus/"
  [6]  (OLDPLACE)  "c:\ado/"
```

- Download the "Course Stata ADO files" zip-file from the Courseplus GENERAL INFORMATION folder. Extract the contents of this zip-file to your personal "ado" file directory.

Theoretically, you should now have access to the "autocor" function (among others) whenever you open a Stata session.

To confirm that the "autocor" command will work on your machine, do the following:

- Download the dataset "test_autocor.dta" from the Courseplus Lab1 folder.
- Open this dataset within Stata.
- Type the following: `autocor y time id`

If the autocor command is working properly, you should see the following output:

```
. autocor y time id

             |    time1    time2    time3    time4    time5
-------------+---------------------------------------------
       time1 |   1.0000
       time2 |   0.8292   1.0000
       time3 |   0.7929   0.8561   1.0000
       time4 |   0.7491   0.8144   0.8816   1.0000
       time5 |   0.7087   0.6830   0.7837   0.8485   1.0000

     +----------+
     |      acf |
     |----------|
  1. | .8485768 |
  2. | .7950424 |
  3. | .7147705 |
  4. | .7086982 |
     +----------+
```

If the autocor command does not work for you, please post to the course bulletin board so we can help you set this up.

R Users: You will need to install the "nlme" package. You will use both the "gls" and "ACF" functions from the "nlme" package to estimate the. To test the use of these functions, do the following:

- Open and R session.
- Download and load the "test_autocor.csv" file as a data object called "data": `data = read.table("test_autocor.csv", sep = ",", header = TRUE)`.
- Install the "nlme" package; `install.packages(""nlme"")`
- Load the "nlme" library: `library(""nlme"")`
- Fit a weighted least squares model to the data: `fit <- gls(y~as.factor(time),data)`
- Obtain the estimated autocorrelation function: `ACF(fit,form= ~1 | id)`

The output should be:

```
  lag       ACF
1   0 1.0000000
2   1 0.8411441
3   2 0.8119743
4   3 0.6927181
5   4 0.7373451
```

Lab Exercise:

1. Simulate data for 100 critical illness survivors using the following multivariate normal distribution.

$$\begin{pmatrix} Y_{i0} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 35 \\ 38 \\ 43 \\ 49 \\ 48 \end{pmatrix}, \begin{pmatrix} 100 & 85 & 80 & 72 & 69 \\ 85 & 100 & 85 & 80 & 72 \\ 80 & 85 & 100 & 85 & 80 \\ 72 & 80 & 85 & 100 & 85 \\ 69 & 72 & 80 & 85 & 100 \end{pmatrix} \right)$$

STATA:

```
set seed 12275
matrix m = (35, 38, 43, 49, 48)
matrix sd = (10,10,10,10,10)
matrix C = (1, 0.85, 0.80, 0.72, 0.69, 1, 0.85, 0.80, 0.72, 1, 0.85, 0.80,
1, 0.85, 1)
drawnorm y0 y1 y2 y3 y4, n(100) corr(C) cstorage(upper) means(m) sds(sd)
gen id = _n
```

R:

```
install.packages("mvtnorm")
library(mvtnorm)
set.seed(12275)
mm <- c(35, 38, 43, 49, 48)
C <- matrix(c(1,0.85,0.80,0.72,0.69,0.85,1,0.85,0.80,0.72,0.80,0.85,1,
0.85,0.80,0.72,0.80,0.85,1,0.85,0.69,0.72,0.80,0.85,1),nrow=5)
sigma <- C * 100
y <- rmvnorm(n=100, mean=mm, sigma=sigma)
id <- seq(1,100)
```



```
dat <- as.data.frame(cbind(y,id))
names(dat) <- c("y0","y1","y2","y3","y4","id")
long <- reshape(dat,varying=1:5,idvar="id",direction="long",v.names="y")
```

2. Compute summary statistics (mean and variance) of the SF-36 mental health scores at each follow-up.

   What is the sample mean SF-36 mental health score at 3 months post hospital discharge?

   ANSWER: Using Stata 13, the sample mean is 50.45. Using R (version 3.3.1), the sample mean is 50.390. NOTE: the answers will vary across Stata and R because the random number generators are different. The answers will vary across different versions of R as well.

   What is the sample standard deviation for the SF-36 mental health score at 1 month post hospital discharge?

   ANSWER: The sample standard deviation at 1 month post hospital discharge is 9.12 (Stata 13) or 10.54 (R).

   Using STATA 13

Time	Baseline	Month1	Month2	Month3	Month4
Sample Mean	36.4236	39.239	45.002	50.450	49.951
Sample SD	10.221	9.116	10.450	9.518	9.869

   Using R

Time	Baseline	Month1	Month2	Month3	Month4
Sample Mean	36.153	39.327	44.523	50.390	48.939
Sample SD	10.896	10.542	9.492	9.148	8.846

   STATA:

```
summarize
```

   R:

```
colMeans(y)
sqrt(diag(var(y)))
```

3. Compute the sample correlation between all pairwise follow-up assessments.

   What is the sample correlation between the SF-36 mental health scores measured at hospital discharge and 4 months post hospital discharge?

   ANSWER: The sample correlation between the SF-36 mental health scores measured at hospital discharge and 4 months post hospital discharge is 0.71 (Stata 13) or 0.74 (R).

	ρ_{01}	ρ_{02}	ρ_{03}	ρ_{04}	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}
STATA	0.8292	0.7929	0.7491	0.7097	0.8561	0.8144	0.6830	0.8816	0.7837	0.8485
R	0.8447	0.8312	0.7204	0.7389	0.8704	0.7875	0.7251	0.8292	0.7900	0.8164

STATA:

```
pwcorr y0 – y4
```

R:

```
cor(y)
```

4. The goal of the study is to estimate the improvement in mental health symptoms among ICU survivors comparing symptoms at 1 to 4 months post discharge to symptoms at hospital discharge. Along with the estimates of improvement, you will provide a confidence interval representing a plausible interval that contains the true improvement.

   Using your simulated data, estimate the improvement with 95% confidence intervals ignoring the longitudinal structure of the data and incorporating the longitudinal structure of the data.

   These computations are easily made with the data in wide format

   STATA:

```
ttest y1=y0, unpaired
ttest y1=y0
```

   R:

```
t.test(dat$y1,dat$y0, paired=FALSE)
t.test(dat$y1,dat$y0, paired=TRUE)
```

   Fill in the table below:

Using STATA 13

Month post hospital discharge	2-sample t-test		Paired t-test		Known correlation
	Mean improvement (95% CI)	CI width	Mean improvement (95% CI)	CI width	
1	2.816 (0.115, 5.517)	5.402	2.816 (1.675, 3.957)	2.282	0.85
2	8.578 (5.696, 11.461)	5.765	8.578 (7.258, 9.899)	2.641	0.80
3	14.027 (11.273, 16.782)	5.509	14.027 (12.634, 15.421)	2.787	0.72
4	13.528 (10.726, 16.330)	5.604	13.528 (12.005, 15.050)	3.045	0.69

Using R

Month post hospital discharge	2-sample t-test		Paired t-test		Known correlation
	Mean improvement (95% CI)	CI width	Mean improvement (95% CI)	CI width	
1	3.175 (0.185, 6.165)	5.980	3.175 (1.988, 4.362)	2.374	0.85
2	8.370 (5.520, 11.220)	5.700	8.370 (7.165, 9.575)	2.410	0.80
3	14.238 (11.431, 17. 043)	5.612	14.238 (12.716, 15.759)	3.043	0.72
4	12.786 (10.018, 15.555)	5.537	12.786 (11.321, 14.252)	2.931	0.69

When using the 2-sample t-test, how does the width of the confidence intervals change as the correlation goes from high (0.85) to moderate (0.69)?

ANSWER: The width of the confidence intervals based on the 2-sample t-test for the four measures of improvement are roughly the same. This is expected as the confidence intervals based on the 2-sample t-test computes the standard error of the difference in the two sample means using ONLY the variance in the outcomes at hospital discharge and the post-discharge months (1 – 4) AND the variance is roughly constant over time (population standard deviation = 10).

When using the paired t-test, how does the width of the confidence intervals change as the correlation goes from high (0.85) to moderate (0.69)?

ANSWER: The width of the confidence intervals based on the paired t-test increase as the time from hospital discharge increases. This is to be expected! The standard error for the difference in the two sample means for the paired t-test incorporates the variance of the individual sample means but also the correlation between the observations. As the time between hospital discharge and the post discharge observations increases the correlation decreases, therefore, the standard error will increase over time creating wider confidence intervals.

Based on the true population parameters, can you justify the patterns you see in the confidence interval width comparing the 2-sample t-test to the paired t-test procedures?

ANSWER: For the confidence intervals based on the two-sample t-test, the width is roughly

$$2 \times 1.96 \times \sqrt{\frac{10^2}{100} + \frac{10^2}{100}} = 5.54$$

For the confidence intervals based on the paired t-test, the width of the confidence interval for the improvement from baseline to month j is roughly

$$2 \times 1.96 \times \sqrt{\frac{10^2}{100} + \frac{10^2}{100} - 2\frac{10 \times 10 \times \rho_{0j}}{100}}$$

The width of the confidence interval will change with j; specifically, the width will increase as ρ_{0j} decreases with j.

5. Now, suppose instead of looking at improvements in the SF-36 mental health scores, you want to describe the general trends in the SF-36 mental health scores over time. Use graphical displays to i) assess trends in the mean and variance of the SF-36 mental health scores over time, and ii) assess heterogeneity within and between individuals over time.

   STATA and R code to generate the figures has been provided in the lab1.do and lab1.R files.

   Do the mean SF-36 mental health scores change roughly linearly over time?

   ANSWER: Roughly linear, sure! But we know from the population means, that the means increase to 3 months post hospital discharge and then the mean decreases at 4 months post hospital discharge.

   What do you think generates the majority of the variation in SF-36 mental health scores? Differences across patients in their scores at hospital discharge? Differences across patients in how fast/slow the SF-36 mental health scores are changing over time? Variation within a patient over time?

   ANSWER: When looking at the spaghetti plot, we see that mostly the subject specific data is parallel; which means the rate of change over time across subjects is similar. There are some subjects whose mental health symptoms fluctuate largely over time but for the most part, the pattern in mental health symptoms within a person over time is consistent. Therefore, the majority of the variation in the data is attributable to differences in patient scores at hospital discharge.

6. Explore the within subject correlation structure by computing the autocorrelation function. First, in your group discuss what defines the "saturated model" for the mean in this example. Fit the "saturated model" for the mean, obtain the residuals and then compute the autocorrelation function.

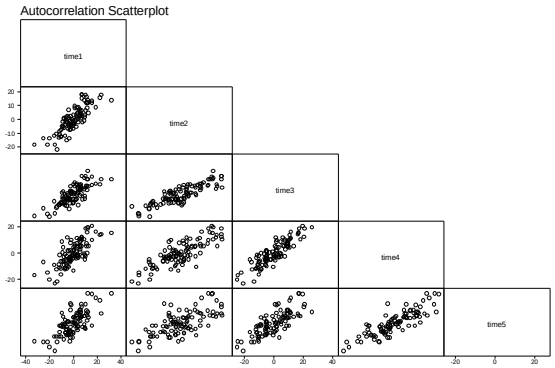
   ANSWER: In this example, time is discrete (hospital discharge and then 1, 2, 3, and 4 months post hospital discharge) and is the primary exposure. Therefore the "saturated model" for the mean is a one-way analysis of variance (ANOVA) with time as the single factor. Alternatively, you can think of this one-way ANOVA as a regression model where we are estimating the mean at each discrete time point: $E(Y_{ij}) = \beta_0 + \beta_1 \times I(time_{ij} = 1) + \beta_2 \times I(time_{ij} = 2) + \beta_3 \times I(time_{ij} = 3) + \beta_4 \times I(time_{ij} = 5)$, where $time_{ij}$ is the follow-up time for survivor *i* at follow-up *j* and $I(time_{ij} = 1)$ defines a dummy variable equal to 1 when $time_{ij} = 1$ and 0 otherwise.

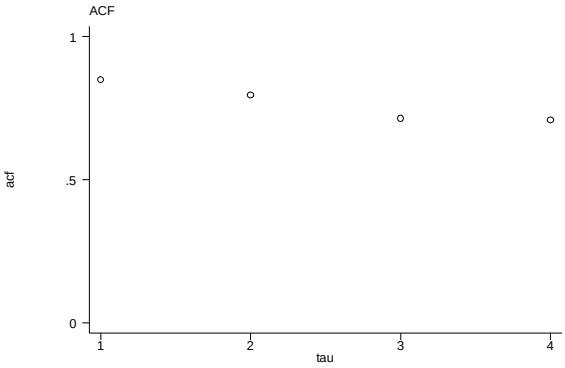
   STATA:

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

```
             |    time1    time2    time3    time4    time5
-------------+--------------------------------------------
       time1 |   1.0000
       time2 |   0.8292   1.0000
       time3 |   0.7929   0.8561   1.0000
       time4 |   0.7491   0.8144   0.8816   1.0000
       time5 |   0.7087   0.6830   0.7837   0.8485   1.0000
```

```
     +----------+
     |      acf |
     |----------|
  1. | .8499038 |
  2. | .7952311 |
  3. | .7153096 |
  4. | .7086982 |
     +----------+
```




```
R:
fit <- gls(y~as.factor(time),long)
ACF(fit,form=~1| id)
  lag       ACF
1   0 1.0000000
2   1 0.8357603
3   2 0.7893926
4   3 0.7233078
5   4 0.7389423
```

In this sample of 100 critical illness survivors, what is the estimated lag 1 correlation? Does this correlation estimate Corr(Y_{i1}, Y_{i2})? Does this correlation estimate any other Corr(Y_{ij}, Y_{ik})?

ANSWER: Using Stata, the estimate is 0.8499 and using R, the estimate is 0.8358. Yes, this is an estimate of Corr(Y_{i1}, Y_{i2}). These values are ALSO estimates of Corr(Y_{i0}, Y_{i1}) , Corr(Y_{i2}, Y_{i3}) and Corr(Y_{i3}, Y_{i4}); i.e. Corr(Y_{ij}, Y_{ik}) such that $|j-k|=1$.

In this sample of 100 critical illness survivors, what is the estimated lag 4 correlation? What values of Corr(Y_{ij}, Y_{ik}) does this correlation estimate?

ANSWER: Using Stata, the estimate is 0.7087 and using R, the estimate is 0.7389. This value corresponds to Corr(Y_{i0}, Y_{i4})

