

Homework # 1

Biostatistics 210

DUE September 27, 2016 (by 10:30am)

0. BACKGROUND

This assignment will give you the opportunity to compare a series of analyses for a simple dataset. The purpose is to give you practice in contrasting analyses and considering their (1) validity (unbiasedness of estimates, size of tests, coverage of confidence intervals), (2) efficiency (standard error of point estimates, power of tests, and width of confidence intervals) and (3) robustness (how methods perform when their key assumptions are incorrect).

The data setting is based on a clinical trial ([AIDS Clinical Trials Group Protocol #5235](#)) of neuroprotective agent for individuals infected with HIV with cognitive decline. The data consists of 100 individuals randomly assigned between placebo and the experimental agent, minocycline. The primary outcome is a “global deficit score” (GDS) which is a normed measure of cognitive impairment. Patients have their GDS measure at enrollment and at week 24.

1. DOWNLOAD

- Download the fake dataset `actg5235.dta` from the course website.

2. THE MODEL

The data that has mean GDS which follows

	Mean Global Deficit Score	
	Baseline	Week 24
Placebo	α	$\alpha + \beta_0$
Minocycline	α	$\alpha + \beta_0 + \beta_1$

based on this. The mean GDS is equal in the two groups at baseline (as expected in a clinical trial) and value is set to be α . The mean GDS is the placebo group at week 24 is $\alpha + \beta_0$ in the placebo group and is $\alpha + \beta_0 + \beta_1$ in the minocycline group at week 24.

The change in GDS in the placebo group from baseline to week 24 is then given by $(\alpha + \beta_0) - \alpha = \beta_0$. The mean difference in GDS at week 24 between minocycline and placebo is

given by $(\alpha + \beta_0 + \beta_1) - (\alpha + \beta_0) = \beta_1$. This is the difference that the study is designed to detect. -- the difference in average GDS scores between the two groups at week 24. One other aspect of this data which we need to consider is the correlation structure. We will see an important aspect of analyzing such data is the magnitude of the correlation, ρ , in the data.

3. ANALYSIS

We will examine four different ways of estimating this difference

1. A simple t-test (compares mean GDS at week 24 between the two groups)
2. A change score t-test (this analysis calculates the change score from baseline to week 24 and then compares those scores between the two groups)
3. A linear regression model for week 24 GDS scores with predictors for treatment and baseline GDS
4. A GEE model which treats GDS values as repeated measures data and models how GDS scores change with time, treatment and a time-by-treatment interaction.

The data was simulated in a way which ensures that all four of these should all estimate the same quantity -- β_1 .

Q1: Among the 4 analysis, which would seem to be the most different in spirit from the other 3? Which, for you, seems to be simplest one to explain?

Read in the data

```
use actg5235.dta
```

The dataset has variables

```
id: Participant ID #
gds1: The Global Deficit Score at baseline
gds2: The Global Deficit Score at week 24
rx: Study Treatment (0=Placebo, 1 = Minocycline)
```

You can quickly get the means in the 2 groups at the two timepoints using the table command

```
table rx, c(mean gds1 mean gds2)
```

You can also get a sense of the data using boxplots

```
graph box gds1 gds2, over(rx) scheme(s1color)
```

Analysis 1: t-test

```
ttest gds2, by(rx)
```

Q2: What is the mean difference given by the t-test, can you identify it as a difference between two cells in the table of means? What is the two-sided p-value for the t-test?

Analysis 2: t-test on change score

First, we create variable diff we create a variable

```
gen diff=gds2-gds1
```

Tabulate the mean change

```
table rx, c(mean diff)
```

Q3: What is the mean change in GDS from baseline to week 24 in each group?

```
ttest diff, by(rx)
```

Q4: What is the mean difference given by the t-test on differences? How does it differ from the one you got from the t-test under analysis 1? Why do they differ? What is the two-sided p-value for the test?

Analysis 3: Linear Regression Model

Compare the effects of treatments using a linear regression model. First, try the model

```
reg gds2 rx
```

then, try

```
reg gds2 rx gds1
```

Q5: What are the treatment effects given by model with and without gds1 as a predictor? Has the coefficient changed? Has the model R^2 changed?

Analysis 4: GEE Model

A final approach is to use all the outcome data by creating a longitudinal model. In such a case, the desired treatment effect is a treatment by time interaction. We need to run a few preliminaries to set up the model.

First switch the data from “wide” to “long”

```
reshape long gds , i(id)
xtset id
gen timepoint=_j-1
```

timepoint will be coded to 0=baseline, 1=week 24. The interaction term is

```
gen inter = timepoint*rx
```

and the GEE command is

```
xtgee gds rx timepoint inter, corr(indep) robust
```

Q6: How do you interpret the results? How do they compare with the other models?

4. SIMULATION

It will be helpful to examine how these 4 analyses compare over many similar datasets. This will give us a sense of the bias, efficiency and power of the techniques.

Download the file

hw1-sim.do

from the course website.

In Stata, go to Window, Do file editor, open new do file

Then Open the File in the do file editor

The simulation allows the following to vary

nsim	the number of simulations to be performed
nobs	the number of people in the trial (1/2 placebo, 1/2 minocycline)
rho	the correlation between global deficit scores at baseline and wk 24
mu1	the mean global deficit score at baseline
mu2	the mean global deficit score at week 24 (placebo group)
sigma1	the SD of global deficit scores at baseline
sigma2	the SD of global deficit scores at week 24
beta:	the effect of minocycline at week 24 (<i>mean difference in GDS betw' minocycline and placebo</i>)

The default values create a dataset which will resemble the one you have analyzed.

Run the simulation using the default values.

The simulation return the following for each simulated dataset

ttest:	treatment effect (estimate of beta) given by t-test in analysis 1
pctest	two-sided p-value given by t-test in analysis 1
power-ttest:	1=if pctest < 0.05, 0 otherwise
paired	treatment effect given by t-test in change score given in analysis 2
ppaired	two-sided p-value given by t-test in analysis 2
power-paired	1=if ppaired < 0.05, 0 otherwise
reg	treatment effect from adjusted regression in analysis 3
preg	two-sided p-value given by regression in analysis 3
power-reg	1=if preg < 0.05, 0 otherwise
gee	treatment effect from gee model in analysis 4
pgee	two-sided p-value given by gee model in analysis 4
power-gee	1=if pgee < 0.05, 0 otherwise

Q7: Are all the method unbiased for estimating the coefficient beta ($=-0.15$)? What are the relative efficiency compared with the GE model? Do any of the methods appear to given an identical estimate of the treatment effect?

Q8: Which method is the most powerful? Which is the least?

Q9: Modify the simulations so $\rho=0.50$. How do your answers in Q7 and Q8 change?

Q10: Modify the simulations so $\rho=0.0$. How do your answers in Q7 and Q8 change?