BIOSTATS 690C - Data Management & Applied Data Analysis with Stata
Fall 2020
Examination 1
DUE Wednesday October 21, 2020
Last Date for Submission for Credit: Wednesday October 28, 2020 (-10 points)

How to do well on this exam:

Aim for publication worthy I will be looking for attention to detail, thoroughness, and clarity. In producing your answers, you will lose points if the work you show me has *not* been edited to remove errors, even if you eventually provide the correct answer. Think of yourself as a consultant delivering a product to a client!

Before you Begin:

Download the following 2 Excel data sets from the course website:

- learndis.xlsx
- gss1000.xlsx

R Users

How to develop your exam submission

Please submit a knitted R Markdown file that you have knitted to PDF To do this, launch R Studio. Open a new R Markdown file. Immediately save it to a saved R Markdown file. Do your work there. I strongly encourage you to create a separate chunk for each question. Submit your PDF to Blackboard.

Stata Users

How to develop your exam submission

Please submit a PDF of a Stata log file that you have saved to Word To do this, launch Stata. At the start of your session, open a new log file (take care to use extension ".log") or append to an existing log file if you are doing your exam work over multiple sessions. Upon completion of your Stata work, exit Stata. Launch Word. Import your saved ".log" file into Word. Edit errors and delete all error messages. Then save your cleaned Word file to PDF. Submit your PDF to Blackboard.

Questions 1-3
learndis.xlsx.
Overview of Learning Disabilities in Children Study

This dataset is a subset of the data from a study conducted by Susan Tomasi and Sharon Weinberg (1999). The data pertain to an analysis of the relationship between academic achievement as measured by math and reading tests and intellectual ability as measured by IQ. The sample is comprised of n=105 children, all classified as learning disabled (LD). As well, all are elementary school aged and from an urban area. One of the variables in this dataset (placemen) refers to the type of placement given to the child, either: part-time resource room placement (these students get resources and additional instruction, both in addition to their regular classroom experience) or self-contained classroom placement (these students are segregated full time).

Reference:
Tomasi S and Weinberg SL (1999). Classifying children as learning disabled: An analysis of current practice in an urban setting. *Learning Disability Quarterly*, **22**, 31-42.

Variable name	Variable Label	Coding/Notes
grade	Grade level	= 1, 2, 3, 4, or 5
gender	Gender	0=male, 1=female
placemen	Type of Placement	0=part-time, 1=full-time segregated
readcomp	Reading comprehension test score	Scores range 0 to 200 (higher is better)
mathcomp	Math comprehension test score	Scores range 0 to 200 (higher is better)
iq	Intellectual ability	Scores range 0 to 200 (higher is better)

__1. (20 points total)

__a) By any means you like, import learndis.xlsx into R Studio or Stata. Tip - While in Excel, take care that the columns are saved in the appropriate formats (numeric or text).

__b) Produce a description of your imported dataset
In Stata the command is **`describe`**
In R, the command is **`str( )`**

__c) Save your imported data to a permanent Stata or R dataset.

__2. (20 points total)

__a) For all variables: Create variable labels

__b) For discrete variables ONLY: Create variable value labels.

__c) For all variables, as needed: Assign missing value codes.

__d) Produce again a description of your imported dataset
In Stata the command is **`describe`**
In R, the command is **`str( )`**

__3. (20 points total)

__a) Create a grouped variable that you name **quart_iq** that has values 1, 2, 3, and 4 according to quartile of value of **iq**.

__b) Label your variable **quart_iq**

__c) Attach variable value labels to the values 1, 2, 3, and 4 of **quart_iq**

Questions 4-5
gss1000.xlsx.
Overview of General Social Survey (GSS)
This data set includes measurements of 42 variables on n=1000. It is a subset of the General Social Survey (GSS), which is an ongoing study of American society conducted by the National Opinion Research Center (NORC).

General Social Survey (GSS)

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop cross-national models of human society; and to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting.

GSS questions include such items as national spending priorities, marijuana use, crime and punishment, race relations, quality of life, and confidence in institutions. Since 1988, the GSS has also collected data on sexual behavior including number of sex partners, frequency of intercourse, extramarital relationships, and sex with prostitutes.

http://www.norc.org/Research/Projects/Pages/general-social-survey.aspx

__4. (20 points total)

The variable **fepol** contains responses to the question "Female not suited for politics" and is coded 1=yes and 0=no. This is potentially confusing since a value of 1 is saying that the respondent believes females are not suited for politics.

__a) Create a more straightforward variable that you name **fepol_yes** and that is a reverse coding of **fepol.**

__b) Label this variable "Females suited for politics"

__c) Assign value labels of "Yes" to the value 1 and "No to the value 0.

__5. (20 points total)

The variable **socbar** contains responses to the question "Spend evening at bar" and is coded 1=never, 2 =once a year, 3=sev times a year, 4=once a month, 5=sev times a mnth, 6=sev times a week, 7=almost daily. There are missing values.

___5a) Create a new variable **socbar4** that is a grouping of the values of **socbar** as follows:

IF socbar =	**THEN code** socbar4 =	**Assign** socbar4 **value label**
missing	1	Unknown
1	2	Never
2 or 3 or 4	3	At most 1x/month
5 or 6 or 7	4	At least several x/month

__5b) Label your new variable.

___5c) Label your new variable values.