BIOSTATS 690C - Data Management & Applied Data Analysis with Stata
Fall 2020
Final Exam
DUE Monday December 7, 2020
Last Date for Submission for Credit: Wednesday December 9, 2020 (-10 points)

For Stata Users
Download the following 2 Stata data sets from the course website:
lung_final.dta
gss_1000.dta

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.

R Users
Download the following 2 R data sets from the course website:
lung_final.Rdata
gss_1000.Rdata

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.

How to do well on this exam

As with the first test, 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!

Questions 1-4
lung_final.dta
lung_final.Rdata

Overview of Lung Function Study (CORD)

The lung function data for this test is a subset of the data from UCLA study of chronic obstructive respiratory disease (CORD). The original study followed over 15,000 persons and obtained measurements of lung function (FVC and FEV1, explained below) at two points in time so that they could investigate the change in lung function in relationship to location of residence, a proxy for exposure to air pollution.

The data used for this test is a subset of that for the first time period only. It represents a sample size of n=150 ***families*** and 5 variables, comprised of study ID plus lung function and associated data for nonsmoking mothers and oldest child.

References:
Detels R., Coulson A, Tashkin D and Rokaw S (1975). Reliability of plethysmography, the single breath test and spirometry in population studies. *Bulletin de Physiopathologie Respiratoire*, **11**, 9-30.

Tashkin DP, Clark VA, Simmons M, Reems C, Coulson AH, Bourque LB, Sayre JW, Detels R and Rokaw S (1984). The UCLA population studies of chronic obstructive respiratory disease. VII. Relationship between parents smoking and children's lung function. *American Review of Respiratory Disease*, **129**, 891-97.

Key:

FVC = Forced Vital Capacity (liters). It is the amount of air that can be forcibly exhaled after taking the deepest breath possible.

Coding Manual

Variable name	Variable Label	Coding/Notes
id	Identification number	1, 2, ….. , 150
area	Area of Residence	1 = burbank 2 = lancaster 3 = long beach 4 = glendora
mfvc	Forced Vital Capacity, mother	Continuous, liters
ocfvc	Forced Vital Capacity, oldest child	Continuous, liters
ocsex	Sex at birth, oldest child	1 = male 2 = female

__1. (10 points total)

One of the objectives of the CORD study was to investigate regional (**area**) differences in lung function, the idea being that exposure to air pollutants might be different depending on where one lives in California and that, possibly, these differences are related to differences in lung function. For this question, the outcome of interest is mom's forced vital capacity (**mfvc**) in each of 2 independent groups: 1) area = 2 ("Lancaster") and 2) area = 4 ("Glendora").

__a) **(3 points)**
By any means you like, produce descriptive statistics for **mfvc** for each of the two independent groups **(area=2** versus **area=4**) ***in ONE table***. **Take care! You will lose points if you produce two separate tables.**

__b) **(3 points)**
By any means you like, test the null hypothesis that the mean FVC for moms is the same in the two areas (Lancaster versus Glendora).

__c) **(4 points)**
In 1-3 sentences, provide a report of your analysis. **Tip: In developing your answer, I encourage you to: 1) tell me what the output says (so I know you're good on that); and then 2) tell me your conclusion (so I know you're good on this, too!).**

__2. (10 points total)

Same setting as question #1. By any means you like, produce a graphical comparison of the distribution of **mfvc** that compares the two independent groups **(area=2** versus **area=4**) ***in ONE graph***.

In constructing your graph, take care:

__a) The y-axis tick marks should be the SAME for both groups
__b) The graph includes a title and any other aesthetics you like!

__3. (10 points total)

Also considered in the CORD study is the possibility that lung function might be influenced by genetics, in particular that a child's lung function might be related to his/her parents' lung function. In this question, consider the paired outcomes defined as [mom's FVC (**mfvc**), oldest child's FVC (**ocfvc**)]. Further, to control for possible sex at birth differences, for this question, consider ONLY those mother-child pairs in which the oldest child is a female sex at birth child (**ocsex=2**)

__a) **(4 points)**
By any means you like, ***in ONE table***, produce descriptive statistics for **mfvc** and **ocfvc.**
Restrict your analysis to the subset of families in which the oldest child is female (**ocsex=2**).

__b) **(3 points)**
By any means you like, test the null hypothesis that the mean FVC for moms (**mfvc**) is the same as that of her oldest child (**ocfvc**). Again, consider only those families in which the oldest child is female sex at birth.

__c) **(3 points)**
In 1-3 sentences, provide a report of your analysis. **Again. In developing your answer, I encourage you to: 1) tell me what the output says; and then 2) tell me your conclusion.**

__4. (10 points total)

Same setting as question #3. By any means you like, produce a graphical summary of the relationship between mom's FVC (**mfvc**) and the FVC of her oldest child (**ocfvc**). As before, consider only those families in which the oldest child is female.

In constructing this graph, take care:

__a) The mother **mfvc** observations and the daughter **ocfvc** observations do **NOT** represent two independent groups. They represent paired data. This means that you should produce a repeated measurements graph.
__b) The graph includes a title and any aesthetics you like.

Questions 5-8
gss_1000.dta
gss_1000.Rdata

Overview of the General Social Survey (GSS)
http://gss.norc.org/About-The-GSS

"For more than four decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States. The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. Altogether the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general as well as the role played by relevant subgroups and to compare the United States to other nations."

The observations and variables in the data set for this exam is a subset of the full 2012 data set. It includes 1000 observations on 7 variables. Take care - Some observations have missing values, which you will have to deal with.

In questions #5 - 8, you will be doing and interpreting some normal theory linear regression analyses of the **happy7.** This outcome variable is "self-reported happiness of the respondent" measured on a 7 point likert scale with 1 = "Completely unhappy" and 7 ="Completely happy". Even though **happy7** is a 7-level variable, we will be doing normal theory linear regression anyway. Specifically, we are interested in exploring the independent and joint significance of 6 predictors: **age, class, children, educ, weekswrk,** and **socfrend.**

Coding Manual

Variable name	Variable Label	Coding/Notes
happy7	Self-reported happiness of respondent	1 -7, numeric with 1 = Completely unhappy 2 = Very unhappy 3 = Fairly unhappy 4 = Neither happy nor unhappy 5 = Fairly happy 6 = Very happy 7 = Completely happy
educ	Highest year of school completed	numeric, 0 - 20
age	Age of respondent	numeric, 18-89
class	Self-reported class	1 = lower class 2 = working class 3 = middle class 4 = upper class
children	Number of children	numeric, 0 – 8 note: 8 = 8 or more
weekswrk	Number of weeks worked in last year	numeric, 0 - 52
socfrend	Spend evenings with friends	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

__5. (10 points total)

Consider first a simple linear regression model in which self-reported happiness (**y=happy7)** is modeled linearly in number of years of education **(x=educ).**

__a) **(2 points)**

Fit the single predictor model. Display the output. Tip – Take a look at question #5b before completing this question, as you will need your output from #5a to answer #5b.

__b) **(3 points)**

Using the output you obtained in #5a, complete the following statements by filling in the blanks. Note: In part "iii", you will need to choose between ***"is"*** and ***"is not"***.

i) **(1 point)**

The % variance of ***happy7*** *that is explained by this model is* _____

ii) **(1 point)**

For each 1 additional year of education completed, happiness is estimated to increase by _____
with associated 95% confidence interval limits _____ *and* _____

iii) **(1 point)**

In this sample of 1000 respondents, years of education completed ***"is"/"is not"***
statistically significantly linearly related to self-reported level of happiness.
(Test of NULL: Slope=0, p-value = _____ *)*

__c) **(5 points)**
By any means you like, produce a ***single graph*** visualization of your simple linear regression that includes both: i) scatter plot of y=**happy7** versus x=**educ**; and ii) Least squares fit of y=**happy7** versus x=**educ.**

__6. (20 points total)

Next, suppose you believe that what really matters *vis a vis* education as a predictor of happiness is the completion of certain milestones, in particular, high school and college. In particular, suppose you want to model education as a predictor of **happy7** using appropriately defined design variables that together represent 3 levels of **educ:**

educ < 12 ("less than high school") *versus*
12 ≤ **educ** < 16 ("high school but not college grad") *versus*
educ ≥ 16 ("college graduate")

__a) **(2 points)**
Create a single grouped measure of educ with 3 levels as defined above that you name **educ_grp**. Produce summary statistics of **educ_grp**.

Value of educ_grp ("label")	Formula using variable educ
1 ("less than high school")	**= 1** IF **educ** < 12
2 ("high school")	= **2** IF 12 ≤ **educ** < 16
3 ("college")	= **3** IF **educ** ≥ 16

__b) **(3 points)**
Create the three 0/1 design variables **educ_less, educ_high,** and **educ_college**. Produce summary statistics for each.

New design variable (0/1)	Formula using variable educ_grp
educ_less	**= 1** IF **educ_grp=1**, **0** otherwise
educ_high	= **1** IF **educ_grp=2**, **0** otherwise
educ_college	= **1** IF **educ_grp=3**, **0** otherwise

__c) **(5 points)**
Fit a linear regression model of **y=happy7** on the design variables you created in question 6b. Use "less than high school" as your referent group. Display your output. Save your predicted values in a new variable that you name **yhat**. You will need these to answer question #6d.

__d) **(5 points)**
By any means you like, ***in ONE table***, produce descriptive statistics for your new variable **yhat**, separately for groups defined by **educ_grp.**

__e) **(5 points)**
By any means you like, produce a ***single graph*** that is a side-by-side plot of the means and 95% CI limits of the predicted values **yhat** for each of the 3 groups of education defined by **educ_grp**

__7. (10 points total)

__a) **(5 points)**
By any means you like, create a dataset called **complete** that is the subset of the observations that are complete on **y=happy7** and all 6 of the potential predictors: **age, class, children, educ, weekswrk,** and **socfrend**.
Show me your code.

__b) **(5 points)**
By any means you like, ***in ONE table***, produce descriptive statistics for **y=happy7** and all 6 of the potential predictors: **age, class, children, educ, weekswrk,** and **socfrend.**

__8. (20 points total)

Suppose that we are particularly interested in the role of education (**educ)** for the prediction of self-reported happiness, both crudely and after controlling for **age, class, children, weekswrk,** and **socfrend.** In this question, you will compare two regression models of **y=happy7**. The first model has just one predictor, **educ**. The second model has 6 predictors: **educ** + **age, class, children, weekswrk,** and **socfrend**.

NOTE: In all of the parts of question 8, restrict your work to the data that are complete on every variable (namely, the dataset **complete** that you created in question 7a.

__a) **(5 points)**
Fit the single predictor model containing the single predictor **educ**. Display your output.
Stata users – use **eststo** to save your model that you name **m1**
R users – save your fit to an object that you name **m1**

__b) **(5 points)**
By any means you like, test the null hypothesis of zero slope on **educ** in the single predictor model that you obtained in questions 8a. In 1-2 sentences, interpret.

__c) **(3 points)**
Fit the six predictor model containing the predictors: **educ, age, class, children, weekswrk,** and **socfrend**.
Display your output.
Stata users – use **eststo** to save your model that you name **m6**
R users – save your fit to an object that you name **m6**

__d) **(3 points)**
Fit the five predictor model containing the predictors: **age, class, children, weekswrk,** and **socfrend**. Display your output.
Stata users – use **eststo** to save your model that you name **m5**
R users – save your fit to an object that you name **m5**

__e) **(5 points)**
By any means you like, test the null hypothesis of zero slope on **educ** after adjustment for **age, class, children, weekswrk,** and **socfrend** in the 6 predictor model containing the 6 predictors **educ** + **age, class, children, weekswrk,** and **socfrend**. In 1-2 sentences, interpret.

__f) **(5 points)**
By any means you like, ***in ONE table***, produce a side-by-side comparison of your three models: **m1, m5,** and **m6.** In 1-2 sentences, what do you conclude? There is no single correct answer here; I'm just looking for you to interpret what you see.