Bianca: Did I understand what you intended on the table?

503/650 2016: Assignment 3 - selecting variables Points missed: \_\_\_\_ out of 25

**Your name: Name of TA:**

In this assignment you will select and clean variables for use in later assignments. For most assignments you need one dependent variable Y and six to eight substantively reasonable independent variables. At least two independent variables (IV) must be continuous and at least two must be binary or categorical variables, called factor variables. For most assignments, you will focus on one continuous variable referred to as C and one factor variable referred to as F. The remaining independent variables are referred to collectively as X.

|  |  |  |
| --- | --- | --- |
| *Type of model (* Stata command *)* |  | *Type of dependent variable (Y)* |
| LRM: Linear regression (regress) |  | Continuous |
| BRM: Binary regression (logit) |  | Binary (0/1) |
| ORM: Ordered regression (ologit) NRM: Nominal regression (mlogit) |  | Ordinal Ordianal (same use used for ologit) |
| CRM: Count regression (nbreg) |  | Count (range must be 7 or greater) |

**1)** Examine the codebooks for the datasets available for class. Codebooks are on Box. Ideally, use the same dataset and independent variables for all assignments. This will save you time and you will gain insights from one assignment to help you in others. Since count outcomes are less common in the class datasets, you might need to use a different dataset for the count models such as Add Health or the science data.

**2)** Select your first and second choice for each type of dependent variable and the associated IVs. You can use different transformations of a variable to create different variables for different levels of measurement, such as years of schooling (continuous) converted to degree status (ordinal or binary). (You can use your first choice DV in assignments unless a TA suggest it will be problematic). Fill out the table on the next page summarizing the variables you are proposing.

Table 1: Summary of variables you propose to use for assignments

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Dependent Variable #1** | **Dependent Variable #2** | **Continuous IV #1** | **Continuous IV #2** | **Factor  IV #1** | **Factor  IV #2** | **Four more IVs** | **Dataset** |
| LRM |  |  |  |  |  |  |  |  |
| BRM |  |  |  |  |  |  |  |  |
| ORM |  |  |  |  |  |  |  |  |
| CRM |  |  |  |  |  |  |  |  |

**3)** Run a clearly documented do-file that examines the variables that you plan to use. In this do-file, you will be cleaning variables and verifying any transformations. A sample do-file is included below. The do-file should have four sections labeled #1 LRM; #2 BRM; #3 ORM; #4 CRM. Each section should have five parts:

Part 1: Describe the variables you will be using along with the level of measurement. Be particularly careful in justifying the level of measurement for your outcome variables.

Part 2: Drop missing data. Use listwise deletion

Part 3: Verify that you transformed the data correctly. Tabulations of the categorical variables to verify they are correct.

Part 5: The results of codebook, compact for all variables.

Part 6: Preliminary regression using the command in the table *using your first choice* of Y (you can but do not have to run the model with your second choice DV). For each model, confirm that C and F are statistically significant (p<0.05). See instruction in 5) for details on running the regressions.

*If the same IVs are used for each outcome, only show Parts 2 and 3 for these variables in the Section for LRM.* That is, you don’t need to show the same variable is correct in each section.

**4)** Turn in this assignment sheet with your name and your log file. The log file **must** be printed in Courier New, 9 pt. font, double sided, no line wrapping, and stapled. **You must post your work in Box.**

**5)** Use the following code to run your regression models. Y is the outcome chosen for each model type. C and F are your continuous and factor variables. X is a set of 4 to 6 other independent variables. Include output that demonstrates that your C and F variables are statistically significant for each model. Your C and F variables can change between assignments. If they do, indicate the change in Part 1 of the appropriate section.

|  |  |  |
| --- | --- | --- |
| *Type of model* |  | *Stata code* |
| LRM: Linear regression models |  | reg Y c.C i.F X |
| BRM: Binary regression models |  | logit Y c.C i.F X |
| ORM: Ordered regression models |  | ologit Y c.C i.F X |
| CRM: Count regression models |  | nbreg Y c.C i.F X |

# When picking variables, keep in mind

**1)** At least one continuous and one factor variable must be statistically significant for each type of model. Choose variables that you expect will have significant effects.

**2)** You can recode a variable from one level of measurement to another. For example, years of schooling could be recoded into a binary variable for having completed high school. When you recode or transform a variable, you **must** create a new variable, not change the values of an existing variable. **Verify that you have transformed the data correctly**.

**3)** Continuous variables must have a clear metric and **cannot be ordinal**. If a variable has a natural metric such as years or dollars, use that metric unless you make a convincing case for some other metric. You might need to recode a variable into more meaningful units. Continuous variables must be in a metric that allows you to see the magnitude of the effect. For example, if you code income in $1, your coefficient might be .000. In this case, you could recode the variable to units of $1000.

**4)** Factor IVs must be coded as whole numbers (1 and 2 are fine, 1.5 and 2.5 are not). You can recode a continuous variable into several groups or recode an existing factor variable. The categories should make substantive sense (e.g., don’t have an age category ranging from 11 – 68). Categorical variables **must have value labels**.

**5)** The ordinal dependent variable must have at least four categories and less than eight.

**6)** Counts with a limited range or that are highly concentrated on one or two values generally will **not** work for the CRM. For example, number of children should not be used. The Add Health and science data sets have many count variables. It is a good idea to find a count variable with a range of at least 7.

**7)** If your variables are similar to those used by someone else, you need to choose new variables. If you have questions, ask.

**An example do-file for this assignment is included on the next page. Note: *do not copy and paste the example below into your do-file*. The formatting from word will cause chaos. Instead, start with the template do-files provided for the course.**

capture log close

log using tdmize-s503-a03-choosing\_variables, replace text

clear all

macro drop \_all

version 14

set linesize 80

set scheme s1manual

// Stat 503 - Categorical Data Analysis

// Assignment #: A03 – Choosing Variables

local who "trent mize"

local pgm "tdmize-s503-a03 "

local dte "2015-08-14"

local tag "`pgm'.do `who' `dte'"

di "Tag: `tag'"

use cda-addhealth4, clear

// #1 LRM: Variables for Linear Regression Model

\* NOTE: only section #2 for BRM is included in this sample do-file

// #2 BRM: Variables for Binary Regression Model

// #2 Part 1: Variables and Level of Measurement

// Proposed Variables \*

//

// varname vartype details

// --------------------------------------------------------------------

// Y1: dadgrades Binary: no recoding needed

// Y2: dadsportBinary: no recoding needed

// C: depress Continuous: no recoding needed

// I: dadedcat Categorical: recode into 3 categories

// X1: white Categorical: no recoding needed

// X2: male Categorical: no recoding needed

// X3: age Continuous: no recoding needed

// X4: hrstv Continuous: recode, top code hours at 70

// BRMinsamp 1 if case is in BRM sample, 0 if case has missing

// values or is otherwise not in BRM sample

// --------------------------------------------------------------------

// #2 Part 2: Recode variables (not needed if recoding done in prior section)

// I: dadedcat, categorical, recoded into 3 categories

\* recode, generate new variable

recode dadeduc (1/6=0) (7/8=1) (9=2) (10/12=.), gen(dadedcat)

label var dadedcat "Father's education group"

label define dadedcatL 0 "HSorLess" 1 "College" 2 "MTcollege"

label values dadedcat dadedcatL

\* confirm tranformation

tab dadeduc dadedcat, miss

// X4: hrtv, continuous, recode to 70 as maximum hours

\* generate new variable

gen hrstvtop = hrstv

label var hrstvtop "Hours of tv per week (truncated)"

\* recode 71 and larger to 70

recode hrstvtop (71/max=70) if !missing(hrstvtop)

\* confirm recode is correct

list hrstv hrstvtop if hrstv > 70 & !missing(hrstv)

\* confirm other values are the same

list hrstv hrstvtop in 1/10

list hrstv hrstvtop in 200/210

// BRMinsamp, binary, 1 if case is in BRM sample

\* count # cases with given number of missing values

egen BRMmissnum = rowmiss(dadgrades dadsport depress ///

dadedcat white male age hrstv)

tab BRMmissnum, miss

\* generate dummy variable indicating no missing values for BRM variables

gen BRMinsamp = BRMmissnum == 0

label var BRMinsamp "In BRM sample"

label define BRMinsampL 0 "Not in sample" 1 "In BRM sample"

label values BRMinsamp BRMinsampL

// #2 Part 3: Tabulation of categorical variables

tab1 dadgrades dadsport dadedcat white male if BRMinsamp==1, miss

// #2 Part 4: Summarize variables

sum dadgrades dadsport depress ibn.dadedcat ///

white male age hrstv if BRMinsamp==1

// #2 Part 5: confirm that C and F are significant

logit dadgrades c.depress i.dadedcat i.white i.male c.age##c.age c.hrstv ///

if BRMinsamp==1, nolog

\*! NOTE: underline significance of C and F variables in your output to confirm

\*! statistical significance at p<0.05

// #3 - ORM: Variables for Ordinal Regression Model

\* only section #2 for BRM is included in this sample do-file

// #4 - CRM: Variables for Count Regression Model

\* only section #2 for BRM is included in this sample do-file

// save data, exit

compress

saveold tdmize-addhealth-clean, replace

label data "`who' cleaned addhealth data `dte'"

log close

exit

CDAiu2016-a03-variables-2016-07-26v2.docx