SOCI 600–601
Introduction to Sociological Data Analysis
Fall 2023

ASSIGNMENT 2
Due by October 17, 2023 (Tuesday) at 11:59pm
Percent of final grade: 14%

ASSIGNMENT 3
Due by November 02, 2023 (Thursday) at 11:59pm
Percent of final grade: 18%

ASSIGNMENT 4
Due by December 07, 2023 (Thursday) at 11:59pm
Percent of final grade: 22%

Instructor information

Ernesto F. L. Amaral, Associate Professor, Department of Sociology
Office: Liberal Arts Social Sciences Building (LASB) 320
Phone: (979)845–9706
Email: amaral@tamu.edu
Course website: http://www.ernestoamaral.com/soci600-23fall.html

Submission

These assignments should be submitted through Turnitin within Canvas. Turnitin is an online database system designed to help instructors **detect plagiarism**, track citations, facilitate peer reviews, and provide paperless grading markup in written assignments. Students should develop this assignment **individually**.

Purpose

The purposes of these assignments are for students to select a research topic, investigate microdata from the American Community Survey with the Stata software package, select and transform dependent and independent variables, examine these variables with the various statistical methods discussed in the course, and organize all this analysis in a final paper draft containing all sections of an academic quantitative paper.

General information

Your grade for these assignments will be determined by the use of **several statistical tools** with a focus on the **quality of your analysis**, and the elaboration of **coherent interpretations**. The accuracy of the formatting of your tables and graphs will also be evaluated. The Stata codes used for this assignment (do-file) should be included at the end of the document as an appendix.

When interpreting tables and graphs, write in plain English, as if you were reporting results in a newspaper. These assignments should be seen as a document that tells a **coherent story about a subject**. Thus, it is important to think wisely about selecting variables for your analysis. You should also make clear that you are estimating characteristics of the Texas population with the American Community Survey (ACS).

You should aim to have a document for all assignments with the following sections: (1) introduction; (2) background; (3) data and methods; (4) results; (5) final considerations; and (6) references. You should place possible tables and figures at the end of the document (after the list of references). Each table and each figure should appear in a separated table. This format makes it is easier to count the number of words you have written. See examples of how to place tables and figures at the end of the document, as well as of how to cite them throughout the document on this link (http://www.ernestoamaral.com/drafts.html).

The documents should be on US Letter paper size, one-inch margins, Arial font, size 11, 1.5 line spacing. Font size within tables can have a smaller size, such as size 9 for numbers and text within tables and size 8 for table footnotes.

Students should take advantage of **regular classes** and **office hours** to clarify any questions with the professor and/or the teaching assistant. The days and time of office hours are listed in the syllabus and course website.

Suggestion of research topic

The research topic for these assignments do not necessarily have to be novel. The intention if for you to analyze microdata and organize the analysis in a final document. If you do not have an idea for a research topic, consider simply analyzing how wage and salary income (dependent variable) is associated with several independent variables (e.g., sex, race/ethnicity, age, educational attainment, marital status, migration status). You can estimate several crosstabulations between variables, tests of statistical significance of income by categories of the independent variables, generate tables, figures, and models for different years, and include interactions between independent variables to explain associations with income.

Assignment 2 (paper outline)

For the paper outline, you should provide an overall **structure of the paper**. You can organize the document in bullet points and/or in short paragraphs. You should highlight the research question, objective, and main hypotheses of your study. You should provide an overall idea of **topics of literature** that you will review. The portion related to the background (literature review) would benefit if you list subtopics that will be reviewed.

You should aim to write around **4 to 5 pages (2,000 to 2,500 words)**, excluding references, tables, figures, and Stata do-file.

You should explain the data and methods you will use. You should **utilize at least the 2021 ACS** throughout your analysis.

You should list some **preliminary results** (e.g., frequency tables, figures, crosstabulations, confidence intervals, tests of significance of main variables by sub-groups of your population). For the results section, you should select an interval-ratio variable to serve as the dependent variable. You should also select at least three independent variables to measure associations with the dependent variable. You should use the appropriate techniques to analyze your variables, depending on their levels of measurement, as explained in lectures and in the section of this assignment about examples of analyses based on course material.

You should include at least 3 illustrations (tables and/or figures) in total:

- At least one illustration for each association between your dependent variable and each of your independent variables.
- Estimate the mean, standard deviation, standard error, and confidence interval of your dependent variable for each category of your independent variables. In order to accomplish this exercise, you should have independent variables at the nominal- or ordinal-level of measurement.

Assignment 3 (paper draft)

For the first draft, you should organize the document in more **elaborated paragraphs**. You should provide a more detailed **literature review** to explain your hypothesis.

You should aim to write around **8 to 9 pages (4,000 to 4,500 words)**, excluding references, tables, figures, and Stata do-file.

You should explain the data and methods you will use. You should **utilize at least the 2005 and 2021 ACS** throughout your analysis.

You should generate **more elaborated results** to help answering your research question and analyze variations over time. In this assignment, you should provide results for associations between the dependent variable and independent variables. As explained below, you should also estimate associations among independent variables. For the results section, you should select an interval-ratio variable to serve as the dependent variable. You should also select at least six independent variables to measure associations with the dependent variable. You can use the same variables from the previous assignment and add three more independent variables. Each of your six independent variables should appear at least once in the tables listed below for both years.

You should include at least 6 illustrations (tables and/or figures) in total:

1. At least one table with **two sample *t*-tests** with equal variances between your dependent variable (interval-ratio-level variable) and at least one independent variable (dummy variable).
    - You can generate a single table and include these tests for more than one independent variable and your dependent variable.
    - Your table should contain the results for 2005 and 2021. If you have too much information to fit in a single table, you can generate more than one table for this item.

2. At least one table with **two-sample test of proportions** between an independent variable (dummy variable) and another independent variable (dummy variable) for each year.
    - You should select an independent variable that can logically be used as a dependent variable for this exercise. For instance, you might use education as an independent variable to explain income in item #1. For this item #2, you can use sex as an independent variable to explain education.
    - Your table should contain the results for 2005 and 2021. If you have too much information to fit in a single table, you can generate more than one table for this item.

3. At least one table with **one-way analysis of variance (ANOVA)** between your dependent variable (interval-ratio-level variable) and at least one independent variable (variable between three and five categories) for each year.
    - Your table should also include the averages of your dependent variable by categories of the independent variable, not only sum of squares, degrees of freedom, mean of squares, F-test and *p*-value.
    - You can generate a single table and include these tests for more than one independent variable and your dependent variable.
    - Your table should contain the results for 2005 and 2021. If you have too much information to fit in a single table, you can generate more than one table for this item.

4. At least one table distribution of with **chi square test** between an independent variable (variable between two and five categories) and another independent variable (variable between two and five categories) for each year.
    - You should select an independent variable that can logically be used as a dependent variable for this exercise. For instance, you might use education as an independent variable

to explain income in item #1. For this item #4, you can use race/ethnicity as an independent variable to explain education.
- Your table should also include the percentage distributions between the two selected variables, population size for the categories in the columns, sample size for the categories in the columns.
- Your table should contain the results for 2005 and 2021. If you have too much information to fit in a single table, you can generate more than one table for this item.

5. At least one table with **Spearman's rho** between an independent variable (ordinal-level variable with at least five categories) and another independent variable (ordinal-level variable with at least five categories) for each year.
   - You should select an independent variable that can logically be used as a dependent variable for this exercise. For instance, you might use education as an independent variable to explain income in item #1. For this item #5, you can use age group as an independent variable to explain education.
   - Your table should also include the percentage distributions between the two selected variables, population size for the categories in the columns, sample size for the categories in the columns.
   - Your table should contain the results for 2005 and 2021. If you have too much information to fit in a single table, you can generate more than one table for this item.

6. At least one **scatterplot** and at least one table with **Pearson's *r***.
   - Generate at least one **scatterplot** between your dependent variable (interval-ratio-level variable) and at least one independent variable (interval-ratio-level variable) for each year.
     - Place your dependent variable in the vertical axis and independent variables in the horizontal axis.
     - Plot a regression line in the scatterplot.
     - Show the least-squares regression equation for the scatterplot (it can be added to the footnote of the scatterplot).
   - Generate at least one table with **Pearson's r** between your dependent variable (interval-ratio-level variable) and at least one independent variable (interval-ratio-level variable) for each year.
     - These should be the same variables utilized for the scatterplot.
     - You can generate a single table and include Pearson's *r* for more than one independent variable and your dependent variable.
     - Your table should contain the results for 2005 and 2021.
     - If you have too much information to fit in a single table, you can generate more than one table for this item.

Note: A variable can be coded in different ways for the several tests:
- Two categories (dummy variable) for the *t*-test and test of proportion. For example, for education: (1) no college degree; (2) at least college degree.
- Between three and five categories for ANOVA. For example, for education: (1) no high school; (2) high school; (3) some college; (4) at least college degree.
- Between two and five categories for the chi square test. For example, for education: (1) no high school; (2) high school; (3) some college; (4) at least college degree.
- At least five categories for the Spearman's rho test. For example, for education: (1) no high school; (2) high school; (3) some college; (4) college; (5) graduate degree.
- At least ten scores for the Pearson's *r* test. For example, for education, IPUMS provides the "educ" variable with 11 categories.

Assignment 4 (final paper)

For the final paper, you should expand the literature review, explain in more detail your **methodology**, and **polish your data analysis and results section**.

You should aim to write around **16 to 18 pages (8,000 to 9,000 words)** in total, not counting references and tables.

You should explain the data and methods you will use. You should **utilize at least the 2005 and 2021 ACS** throughout your analysis.

In this assignment, you should have a consistent group of descriptive results, as well as more elaborated regression results (tables with progressive regression models, standardized regression coefficients, analysis of multicollinearity, figures with predicted values of the dependent variable, figures with residual analysis).

You should also provide a more **cohesive introduction and final considerations** to connect your whole analysis. For the results section, you should select an interval-ratio variable to serve as the dependent variable. You should also select at least six independent variables to measure associations with the dependent variable. You can use the same variables from the previous assignment.

You should include at least 12 illustrations (tables and/or figures) in total:

- At least one illustration for each association between your dependent variable and each of your independent variables (at least six illustrations).
- At least one table with at least three progressive models and standardized coefficients for the full model (at least one table).
- At least one table with variance inflation factors (VIF) for the full model (at least one table).
- At least three figures with predicted values of your dependent variable by scores of independent variables (at least three figures).
- At least one figure with regression residuals by predicted values of your dependent variable (at least one figure).

Examples of analyses based on course material

Estimating means and proportions (Healey, chapter 7)

Estimating means
– Estimate means (interval-ratio or ordinal-level variable with at least three scores), standard errors, sample size, and construct 95% confidence intervals for the mean of your variable for each year

Estimating proportions
– Estimate proportions (nominal- or ordinal-level variable), standard errors, sample size, and construct 95% confidence intervals for the proportions of each category of your variable for each year

Analyzing results
– Include in your analysis a brief explanation of the role of these concepts and terms in the estimation: sample, population, statistic, parameter, equal probability of selection method (EPSEM), representative, and confidence level. This can be part of the overall data and methods paragraph.

Two-sample *t*-test & Two-sample test of proportions (Healey, chapter 9)

Two-sample *t*-test
– Estimate two-sample *t*-tests with equal variances between your dependent variable (interval-ratio-level variable) and independent variable (dummy variable) for each year

Two-sample test of proportions
– Estimate two-sample test of proportions between your dependent variable (dummy variable) and independent variable (dummy variable) for each year

Analyzing results
– Report and explain the results of the tests. At a minimum, your analysis should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic, the results of the test, and the confidence level.

Analysis of variance & Chi square (Healey, chapters 10–11)

Analysis of variance (Healey, chapter 10)
– Estimate one-way analysis of variance between your dependent variable (interval-ratio-level variable) and independent variable (variable between three and five categories) for each year

Chi square (Healey, chapter 11)
– Estimate the chi square tests between your dependent variable (variable between two and five categories) and independent variable (variable between two and five categories) for each year

Analyzing results
– Report and explain the results of the tests. At a minimum, your analysis should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic, the results of the test, the degrees of freedom, and the confidence level

Dependent variable at the ordinal level of measurement (Healey, chapter 12)

Dependent variable at the ordinal level of measurement
– Generate bivariate tables between a dependent variable (ordinal-level variable with at least five categories) and an independent variable (ordinal-level variable with at least five categories) for each year.

– Place your dependent variable in the rows and independent variables in the columns of the cross tabulations. Show column percentages and column absolute totals for every table.

– Estimate chi square, Cramer's V, Lambda, Gamma, and Spearman's rho for every table you request.

Analyzing results
– Report and explain the results of these correlations. For each combination of variables, report the appropriate measures of associations and tests of significance. Analyze the statistical significance, strength (i.e., importance, magnitude), and pattern/direction of the associations between the dependent variable and independent variables.

Dependent variable at the interval-ratio level of measurement (Healey, chapter 13)

Generate scatterplots
– Generate a scatterplot between a dependent variable (interval-ratio-level variable) and an independent variable (interval-ratio-level variable) for each year.

– Place your dependent variable in the vertical axis and independent variables in the horizontal axis.

– Plot a regression line in all scatterplots.

– Show the least-squares regression equation for each scatterplot (it can be added to the footnote of the scatterplot).

Estimate Pearson's *r*
– Estimate Pearson's *r*, test of significance for Pearson's *r*, and coefficient of determination (r^2) for each association of a dependent variable (interval-ratio-level variable) and independent variables (interval-ratio-level variables) for each year.

Analyzing results
– Report and explain the results of these correlations. Answer the following questions with scatterplots:
1) is there an association between the variables?
2) How strong is the association (i.e., importance, magnitude)?
3) What is the pattern/direction of the association?

– Furthermore, use scatterplots to check for linearity.

– Interpret the estimated Pearson's *r*, tests of significance for Pearson's *r*, and coefficients of determination (r^2).

Regression models (Healey, chapter 15; Treiman, chapters 5, 6, 7, 9, 10)

Estimate ordinary least squares regressions (linear regressions)
– Estimate least-squares multiple regression model for each year.

– Make appropriate transformation of dependent variable to have it closer to a normal distribution (e.g., natural logarithm of income).

– Generate dummy variables for independent variables to estimate more meaningful results.

– Generate interaction between independent variables when appropriate. You can generate a series of dummy variables combining values of two or more variables (e.g., indicators of age-education groups).

– Estimate progressive models (e.g., add one variable or one group of variables at a time).

– Report estimated coefficients, robust standard errors (between parentheses), statistical significance (with asterisks), coefficient of multiple determination (R^2), and sample size for each model.

– Include the standardized regression coefficients (i.e., standardized partial slopes, beta-weights) in a separate column for the full model.

– Estimate variance inflation factor (VIF) to verify possible issues of multicollinearity.

– Generate figures with predicted values of the dependent variable by scores of independent variables.

– Generate figures to conduct residual analysis: Scatterplots of residuals by other variables (original dependent variable, predicted dependent variable, and independent variables).

Analyzing results
– Report and explain the results of these models.

– Analyze the statistical significance, strength (i.e., importance, magnitude), and pattern/direction of the associations between the dependent variable and independent variables.

– Interpret standardized coefficients.

– Interpret multicollinearity (based on VIF).

– Explain the overall fit of each model (based on R^2).

– Explain which model is the most parsimonious, i.e., it provides more understanding of variation of the dependent variable with fewer independent variables (based on adjusted R^2).

– Explain which independent variable has the strongest impact on the dependent variable (based on standardized regression coefficients).

Other considerations

1) The American Community Survey (ACS) microdata is available on the course website, as well as from the IPUMS website (https://usa.ipums.org/usa-action/samples).

2) See the ACS codebook in the IPUMS website (https://usa.ipums.org/usa-action/variables/group) for a list of available variables. You should generate new variables to recode original ones if appropriate, as we performed in class.

3) You should avoid including tables and figures in your assignment that do not enhance (or are not related to) your analyses. You should analyze all tables and figures included in your assignment.

4) If reporting missing cases, do not include them in the total of the tables. Preferably, report missing cases in a row below the total. There are three different types of missing values in ACS. For instance, there are respondents who are not asked to answer a specific question, so the variable for those cases are assigned as not applicable (N/A) or observations not in universe (NIU). In other cases, respondents might not have provided information, so the variable has a missing case. In most cases, it would be better to differentiate between these types of missing cases by including one row for each of them at the bottom of the table.

5) You should utilize appropriate formatting for your tables and graphs. This file has several examples of how to correctly format tables and graphs (http://www.ernestoamaral.com/docs/soci600-23fall/Examples_tab_fig.pdf). There are also some papers (http://www.ernestoamaral.com/papers.html) and drafts (http://www.ernestoamaral.com/drafts.html) on my website, which can help you with the correct format for tables and graphs.

6) You can copy tables from Stata to Word (highlight table, right click, and select “Copy table as HTML”) in order to format them. You can also copy tables from Stata to Excel (highlight table, right click, and select “Copy table” or “Copy table as HTML”), format them, and copy to Word. I suggest copying tables from Excel to Word in an editable format, instead of pasting as figures.

7) If it is complicated to generate all graphs in Stata, you can copy tables from Stata to Excel to generate graphs. There are several examples of how to generate graphs in Excel on the course website (http://www.ernestoamaral.com/docs/soci600-23fall/Excel_charts.zip).

8) Several variables and a large amount of information can be organized in a single table in a clear and objective manner. For example, look at Table 1 (frequency distributions), Table 2 (percentage of one variable by categories of other variables), and Tables 3, 4, and 5 (statistical regressions) in the paper about characterization of fertility levels in Brazil (http://doi.org/10.17605/OSF.IO/8FRJ4). You can also see Table 1 (frequency distributions), Table 2 (rates of one variable by categories of other variables), and Tables 3 and 4 (statistical regressions) in the paper about rising cesarean section rates in Brazil (http://doi.org/10.17605/OSF.IO/QFHXE). There are also other papers on my website that provide additional examples.

9) You can illustrate descriptive statistics using graphs, instead of tables. For example, look at Figures 2 and 3 in the paper about the growth of Protestantism in Brazil (http://doi.org/10.17605/OSF.IO/C5P2A).

10) You should perform the data analysis with the statistical software Stata. The codes generated in this software (do-file) must be included at the end of the assignment as an appendix.

11) You should use the person weight (“perwt”) on your analysis.

12) You can simply use the survey weight if you are estimating only frequency distributions and measures of central tendency (e.g., mean, median). However, you need to utilize the complex survey design (“svyset” and “svy”) if you are estimating measures of dispersion (e.g., standard deviation, standard error), margins of error, confidence intervals, and statistical significance (e.g., *t*-test, *p*-value).

13) The command “summarize” provides descriptive statistics for the sample. It does not provide inferential statistics for the population. You would have to indicate the complex survey design with the command “svyset” to get the standard error of the estimate of the population mean. The command “svy: mean” (followed by “estat sd”) provides an estimate of the population mean and an estimate of its standard deviation. When computing the standard error, consider the effect of clustering and stratification, as well as the effect of sampling weights (i.e. complex survey design). However, the clustering and stratification do not affect the point estimate of the mean. Thus, if you are interested only in the point estimate (e.g. mean, median), you can use “summarize” with “aweights” since it gives the same weighted mean as “svy: mean.” For quantiles, “summarize” with “aweights,” as well as “pctile” with “aweights” or “pweights,” all give the same answers. If you use “summarize” with “aweight” (not considering the complex survey design), this strategy assumes a simple random sample, in which: (1) an estimate of the population mean is the sample mean; and (2) an estimate of the population standard deviation is the sample standard deviation. By not indicating complex survey design variables, Stata will assume a simple random sample and underestimate standard errors. You should explain this limitation in interpreting the weighted standard deviation, when not indicating the complex survey design (https://www.stata.com/support/faqs/statistics/weights-and-summary-statistics).

Considerations for regression models (not applicable for all assignments)

1) Regression models should be estimated considering the ACS complex survey design.

2) It is possible to illustrate several regression models in a single table. Remember to include estimated coefficients, robust standard errors (between parentheses), and statistical significance (with asterisks). You can also illustrate the standardized regression coefficients in separated columns. Use the command “outreg2” to transfer the regression models from Stata to Word. If the “outreg2” command is not available in your Stata software, you can install it from this file (http://www.ernestoamaral.com/docs/soci600-23fall/Modules.zip).