**PM511a: Final Project**

**Spring 2021**

Assigned: Apr 23, 2021; 1:00 pm

Due: May 6, 2021; 11:59 pm

This final project will test your skills with data management, data analysis, and summarization. From blackboard, please download and use the datasets specific to your PM511a student number. Your datasets are unique and unlike the datasets of your classmates.

There are 100 points possible in this exam (4 questions), plus 10 possible extra credit points (2 extra credit questions). Point values for each question are indicated in parentheses. You should turn in a separate electronic document for each question, organized in the following order: written summary/ tables/figures (restricted length) and an appendix with your SAS program, followed by relevant output (unrestricted length). You should therefore turn in a total of 4 electronic documents or 6 electronic documents if you answer both of the extra credit questions. **All written summaries should use Arial 11pt font, with single spacing and 1-inch margins on all sides.** Tables/output can be presented in a fixed width font, like Courier New, if necessary.

All documents must be in **MS Word** (except the Certification of Academic Integrity). Each file name should indicate the question (Part A, B, or extra credit (EC)) and question number (1 or 2), therefore, please use the following naming convention: **Lastname\_firstname\_PartAQ1.docx**.

All work must be your own. Unlike the homework assignments, you may **not** discuss strategies with students or anyone else, including tutors or the TAs. Dr. Choudhury will answer any questions of clarification you might have. Evidence that you have used outside help in any way will result in zero credit for your final project, a grade of F for the class, and a report describing the incident submitted to your department chair and to the graduate school. Evidence of shared work between two or more students will result in the same penalties for all involved students. Your project will not be graded unless you have signed and turned in the signed Certification of Academic Integrity form.

Your project is due by 11:59 p.m. on Friday, May 6 on Blackboard. Late projects will not be accepted. Projects may be turned in early.

(continued on next page)

**PART A: DATA**

CHARGE (Childhood Autism Risks from Genetics and the Environment) is a case control study that was launched in 2003 as the first comprehensive study of environmental causes and risk factors for autism and developmental delay. The CHARGE study recognizes that both genes and non-inherited factors contribute to autism, developmental delays, and children’s behaviors.

The *Mullen Scales of Early Learning* (*MSEL*) is a standardized assessment that is commonly used in clinical psychology as a developmental measure of cognitive development. An early learning composite score can be derived from the 5 subscales: fine motor, visual reception, receptive language, and expressive language scales. For young children this early learning composite score is considered equivalent to a more traditional “IQ” score or a developmental standard score. Each subscale is standardized to calculate a standard score, percentile and age-equivalent score. Usually MSEL scores range from 0-162. MSEL was administered to all children in the CHARGE study and the early learning composite score was calculated for each child.

A modified portion of the data has been made available to you for this project and is stored in the dataset “**charge\_XXX.sas7bdat**” where XXX is your assigned student number for this Amcourse. You are not permitted to use this data for anything other than this exam.

|  |  |  |
| --- | --- | --- |
| Variable | Description |  |
| subject\_id | Subject ID\*\* |  |
| msel\_elcs | MSEL early learning composite score |  |
| folicacid\_t1 | Maternal folic acid (𝜇g) intake during the first trimester |  |
| mthfr677 | Child's MTHFR677 Genotype | 0: CC or CT genotype  1: TT genotype |
| prepregbmi\_cat | Mother's pre-pregnancy BMI | 1: < 18.5  2: 18.5 to <25.0  3: 25.0 to <30.0  4: 30.0+ |
| yob | Child's year of birth |  |
| child\_male | Child's sex | 0: Female  1: Male |
| agemomyrs | Mother's age at birth (years) |  |
| maxedu\_bs | Maximum education in the home | 0: Some college or less  1: Bachelor's degree or higher |

(\*\***Note:** In this dataset, **each** observation is from a **different** person. This particular ID variable is a de-identified mock ID. There might be some redundancy/ duplication in this ID variable and hence, it should be ignored.)

**PART A: Questions**

1. (15 points) Using **charge\_XXX.sas7bdat**, perform the necessary analyses to investigate the pairwise relationships of the MSEL early learning composite score with each of the variables listed above. Create a table with appropriate descriptive statistics of each variable as well as one or more statistics that the informatively summarizes that variable’s pairwise association with MSEL early learning composite score (include the accompanying *p*-value for testing the null hypothesis of no association). Following Table 1, describe the methods you used **(<1/2 page)** and summarize your results **(<1/2 page)** in language appropriate for a medical journal. Report all values to 3 decimal places (truncate very small p-values to <0.001).

Table

Description automatically generated

1. (40 points) Maternal intake of folic acid early in pregnancy has been found to be associated with Autism Spectrum Disorder and other developmental disorders. Researchers are curious if this intake affects IQ during the developmental stages of childhood. Investigate the relationship between MSEL early learning composite score and folic acid intake during the first trimester of pregnancy. You should consider all given variables in your analyses. In the style of a journal article, write a “Methods” section describing your analytic approach, and summarize your findings in a “Results” section. Combined, your “Methods” and “Results” sections **must not exceed 2 pages** (excluding tables and figures). You can assume that the audience for your written summary is familiar with all aspects of data analysis covered in this course. You may include tables and/or figures after your write-up. Please label them clearly and sequentially and refer to them within your written summary. Tables must be created in Word and only relevant SAS-generated figures should be included. **Do not include any SAS output or code in your written sections**. Please append code and relevant output AFTER your written sections and AFTER any tables/figures.

**Extra Credit 1 (5 points – all or nothing)**: A pilot study is being created to examine the association between newborn dried blood spot folate and incidence of ASD among mothers that had high folic acid intake (> 800 𝜇g) during the first trimester. The researchers want to only include those mothers whose babies were born in recent years (2004 and later) and that were not obese prior to pregnancy. Write SAS code that randomly selects 15 subjects from this dataset to be included in the new pilot study. (Hint: you have generated random numbers previously in SAS).

Create a brief report that includes your methods (**1/2 page**) and a print-out (screenshot of PROC PRINT is ok) of the following information for only those participants who are **selected** for this study: subject id, folic acid intake, BMI, year of birth, and sex of the baby.

**PART B: DATA**

National Maternal and Infant Health Survey (NMIHS) conducted a survey in 1988 to study factors related to poor pregnancy outcomes. A stratified random sample of 26,355 mothers from 48 states, the District of Columbia, and New York City were selected. These mothers were mailed questionnaires and 18,594 mothers returned the questionnaire. Modified questionnaire responses and data gathered from birth certificates from mothers whose pregnancies were nearly full term (between 37 and 41 weeks) were used to create the datasets for Part B.

|  |  |  |
| --- | --- | --- |
| Variable | Description |  |
| momAge | Mothers age at delivery, years |  |
| educ | Highest Level of Education attained by mother | 0: <high school  1: high school 2: some college 3: college 4: some grad school |
| prenatalClass | Ever taken a prenatal/childbirth class | 0: No  1: Yes, this pregnancy  2: Yes, previous pregnancy only |
| smoking | Smoking before and during pregnancy | 0: non-smoker (before and duringpregnancy)  1: light smoker (< 10 cig/day **before &** **during** pregnancy)  2: moderate smoker (≥10 cig/day **before** pregnancy & <10 cigarettes/day **during** pregnancy)  3: heavy smoker (≥10 cig/day **before** **and during** pregnancy) |
| parity | Total number of children ever born to the mother (including this one) | 1: First child 2: Second child 3. Third or subsequent child |
| ~~wtprepreg~~ | ~~Mother's pre-pregnancy weight (in pounds)~~ |  |
| BMIprepregCat | Mother's pre-pregnancy BMI | 0: underweight  1: normal weight  2: overweight  3: obese |
| wtGain | Weight gain during pregnancy, pounds |  |
| gestWeeks | Duration of pregnancy, weeks |  |
| male | Sex of the infant | 0: Female  1: Male |
| bwt | Birthweight of the infant, grams |  |

**PART B: Questions**

1. (30 points) Use **nmihs88\_train\_XXX.sas7bdat** to predict infant birthweight (bwt) in non-smoking mothers using all available variables. When developing your model(s), you should consider higher orders of these candidate predictor variables (*i.e.*, allow for possible nonlinearity of continuous predictors). Consider interactions with each continuous predictor and the following predictors: male and bmiprepregcat. Summarize your work by writing a brief methods section and a brief results section, as if for a journal manuscript. Combined, your “Methods” and “Results” sections **must not exceed 2 pages** (excluding tables and figures).

**Extra Credit 2 (5 points – all or nothing)**: Validate your model in Question B.1 by applying it to the “holdout” dataset (**nmihs88\_hold\_XXX.sas7bdat)**. Describe your methods, results, and conclusion in a brief **< 1/2 page write-up** (exclusive of any relevant tables/figures).

1. (15 points) Previous studies have pointed to 3rd trimester smoking as the most dangerous trimester (note: NMIHS data do not contain information on smoking by trimester), so women who stop smoking during pregnancy have the potential to improve the health of their babies. You are planning a new study to evaluate the impact of a new smoking intervention program for pregnant women who are heavy smokers at the time they are recruited, in the first trimester. The new intervention will be randomly assigned to 50% of study participants, while the remaining participants will receive standard counselling from their doctor. What is the sample size required to detect an increase of 250 grams in infant birthweight associated with the intervention? Use data from the subset of women who were heavy smokers before and during pregnancy in the NMIHS to calculate any required information. Clearly state your: (a) assumptions/parameters, (b) methods and/or software you used to obtain your estimated sample size and (c) result in a **< 1/2 page write-up** (exclusive of any relevant tables/figures).