Homework 3: Working with messy data
Due 1/30/2023 (50 points)

Reports counts of live births registered in the United States.

There are two versions of the natality file on the chalk site – a 10% file (`bigsample.dta`) and a 1% file (`littlesample.dta`). An updated version of the larger data file may be used for the final presentation (find this at http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm). Unlike the Framingham dataset, there are a lot of messy variables and missing data in these natality datasets.

Use the 1% sample for this assignment. Happily, the "do file" you create to clean the data will also work in the 10% file.

Goals of this homework: To create an analysis data set using a do file and to examine correlates of missingness.

Create analysis data set

Look at `Natality2002.pdf`, which describes the data. There are many variables you will never use in this data set. Note that a few of the variables described there are not included in your data sets.

For now, imagine that from a research perspective you are interested in the effects of maternal age, race and education on risk of preterm birth (gestational age < 37 weeks). Focus on these four variables (*maternal age; maternal education; maternal race; gestational age*) for this data cleaning and re-categorization exercise.

For each of these variables, examine the distribution, note the presence of missing values (if any) and think about re-categorization so that each variable is represented by no more than 5 categories (fewer is adequate for all variable but race). **Remember! Make sure missing data are indicated by "." rather than by a number.**

[A few variables include letters as well as numbers ("ZZZ"); these are defined as string variables. If you want to do any numerical manipulation with these variables, you will need to replace the Zs with numbers or missing indicators (.) and then convert them to a number variable.]

a. Turn in the program (do file and log file) that creates the analytic file via commands for cleaning and recategorizing these variables. Make sure this do-file is annotated (using asterisks at the beginning of the annotation line) so that your goals and rationale are clear. ***(10 points)***

b. Create a table that compares those with and without missing gestational age in terms of the other three variables. Test whether the distribution across categories of race, education and age is different for those with missing gestational age and report a p-value for the tests (use what you learned from statistical method classes). ***(30 points)***

This is typically the format for such a table:

Gestational Age (<37 weeks)

	Complete data	Missing gestational age	p-value
✓ Variable 1 Maternal Age			0.xx
Group a	xxxx	Xx	
Group b	xxxx	Xx	
Group c	xxxx	Xx	
✓ Variable 2 Maternal Education			0.xx
Group a	xxxx	xx	
Group b	xxxx	xx	
Group c	xxxx	xx	
Group d	xxxx	xx	

✓ Variable 3 Maternal Race

c. Describe in a short paragraph what the table shows. ***(10 points)***

Epidemiologic Methods (PBHS 31001, Winter 2024)

Alexandra Chang
01/30/2024

Homework #3

Problem 1.

Solution.

Re-categorization of Each Variable with Hypothesis Testing

(Refer to the STATA .do and .log file for complete steps)

```
. tab Maternal_Age Gestational_Age, chi
```

RECODE of dmage	0=Complete gestation, 1=Missing gestation 0	1	Total
< 15 yrs old	**241**	**8**	**249**
15-25 yrs old	**16,117**	**195**	**16,312**
25-35 yrs old	**19,325**	**165**	**19,490**
35-45 yrs old	**4,153**	**45**	**4,198**
45-55 yrs old	**25**	**0**	**25**
Total	**39,861**	**413**	**40,274**

Pearson chi2(**4**) = **22.8741** Pr = **0.000**

Figure 1: Chi-square Comparing **Age Categories** with Presence of Gestation Data

- **Maternal Age** has been re-organized into **five distinct categories**, each spanning a decade, ranging from under 15 years to 55 years, as indicated in the grouped chart above.
- The output displays relevant statistical information, including degrees of freedom (chi2(4)), a chi-squared value (22.87), and a P-value (0.000). It is important to note that the reported P-value cannot be precisely zero; in Stata, a P-value of 0.000 is essentially P<0.0005, leading to the recommendation of reporting **P<0.001 (Rejection of null).**
- In this context, a small p-value suggests that the distribution of **maternal age varies significantly across different gestational age statuses**.

```
. tab Maternal_Education Gestational_Age, chi
```

1=No education, 2=Elementary, 3=Highschool, 4=College, 5=Not stated	0=Complete gestation, 1=Missing gestation 0	1	Total
1	**77**	**0**	**77**
2	**2,295**	**43**	**2,338**
3	**18,159**	**195**	**18,354**
4	**18,796**	**129**	**18,925**
5	**534**	**46**	**580**
Total	**39,861**	**413**	**40,274**

Pearson chi2(**4**) = **310.8476** Pr = **0.000**

Figure 2: Chi-square Comparing **Education Categories** with Presence of Gestation Data

- **Maternal Education** has been re-organized into **five distinct categories**, spanning from no education, elementary school, high school and college, as indicated in the grouped chart above. Missing values are indicated as group 5.
- The output displays relevant statistical information, including degrees of freedom (chi2(4)), a chi-squared value (310.85), and a P-value (0.000). It is important to note that the reported P-value cannot be precisely zero; in Stata, a P-value of 0.000 is essentially P<0.0005, leading to the recommendation of reporting **P<0.001 (Rejection of null).**
- In this context, a small p-value suggests that the distribution of **maternal education varies significantly across different gestational age statuses**.

```
. tab Maternal_Race Gestational_Age, chi
```

RECODE of mrace	0=Complete gestation, 1=Missing gestation: 0	1	Total
White	31,328	348	31,676
Black	6,020	31	6,051
American Indian	424	4	428
Chinese	329	5	334
Japanese	94	0	94
Hawaiian	73	0	73
Filipino	324	10	334
Asian Indian	312	1	313
Korean	88	1	89
Samoan	16	0	16
Vietnamese	144	5	149
Guamanian	8	0	8
Other Asian/Pacific I	246	4	250
Combined Asian/Pacifi	455	4	459
Total	39,861	413	40,274

Pearson chi2(13) = 43.3044 Pr = 0.000

Figure 3: Chi-square Comparing **Race Categories** with Presence of Gestation Data

- **Maternal Race** has been organized into **14 distinct categories**, as indicated above.
- The output displays relevant statistical information, including degrees of freedom (chi2(13)), a chi-squared value (43.30), and a P-value (0.000). It is important to note that the reported P-value cannot be precisely zero; in Stata, a P-value of 0.000 is essentially P<0.0005, leading to the recommendation of reporting **P<0.001 (Rejection of null).**
- In this context, a small p-value suggests that the distribution of **maternal race varies significantly across different gestational age statuses**.

Problem 2.

Solution.

Summary Table for Descriptive Statistics for Various Variables

(Refer to the STATA .do and .log file for complete steps)

- Please refer to Table 1 on the last page of this HW#3.
- Regarding those with and without gestational age, as indicated by Table 1, each variable representing maternal characteristics (age, education, race) all demonstrate a **statistically significant difference when stratified across gestational categories (P< 0.001)**.
- There is a clear pattern observed in maternal age, where gestational **age peaks within the 25-35 age groups** for both complete and missing gestational age categories, and subsequently **decreases with advancing age**.
- **Complete gestational data** exhibits its highest occurrence within the maternal **college** education group, while the maternal **high school** education group records the highest frequency for **missing gestational data**.
- **Both** complete and missing gestational data show the **highest frequency within the maternal White** and Asian racial groups.

```
. table1, by(Gestational_Age_) vars(Maternal_Age cat\ Maternal_Education cat\ Maternal_Race cat) format(%2.1f) missing
```

Factor	Level	Complete Gestational	Missing Gestational	p-value
N		39861	413	
Maternal_Age	< 15 yrs old	241 (0.6%)	8 (1.9%)	<0.001
	15-25 yrs old	16117 (40.4%)	195 (47.2%)	
	25-35 yrs old	19325 (48.5%)	165 (40.0%)	
	35-45 yrs old	4153 (10.4%)	45 (10.9%)	
	45-55 yrs old	25 (0.1%)	0 (0.0%)	
Maternal_Education: 1=No edu, 2=Elementary, 3=High, 4=College, 5=Missing	1	77 (0.2%)	0 (0.0%)	<0.001
	2	2295 (5.8%)	43 (10.4%)	
	3	18159 (45.6%)	195 (47.2%)	
	4	18796 (47.2%)	129 (31.2%)	
	5	534 (1.3%)	46 (11.1%)	
Maternal_Race	White	31328 (78.6%)	348 (84.3%)	<0.001
	Black	6020 (15.1%)	31 (7.5%)	
	American Indian	424 (1.1%)	4 (1.0%)	
	Chinese	329 (0.8%)	5 (1.2%)	
	Japanese	94 (0.2%)	0 (0.0%)	
	Hawaiian	73 (0.2%)	0 (0.0%)	
	Filipino	324 (0.8%)	10 (2.4%)	
	Asian Indian	312 (0.8%)	1 (0.2%)	
	Korean	88 (0.2%)	1 (0.2%)	
	Samoan	16 (<1%)	0 (0.0%)	
	Vietnamese	144 (0.4%)	5 (1.2%)	
	Guamanian	8 (<1%)	0 (0.0%)	
	Other Asian/Pacific Islander	246 (0.6%)	4 (1.0%)	
	Combined Asian/Pacific Islander	455 (1.1%)	4 (1.0%)	

Figure 4: Descriptive Table for Maternal Characteristics with Presence of Gestational Data.