MSCR 520: Homework 3

Anish Shah

April 8th, 2020

The research question is:

Controlling for gestation, plurality, mother age, and mother white, what is the difference in birth weight between smoking mothers and nonsmoking mothers? Notice that in the research question above, not all the variables in the data set are used. This is because proc mi notoriously involves extremely time-consuming calculations. To make sure that Citrix does not log you off while you are running SAS, I have to take a subset of babies born in 2012 as well as select less than a handful of variables.

GENERAL DIRECTIONS: Whenever you use proc mi to answer the questions below, always use seed=83743. In practice, you can use any seed or use a new seed whenever you call a new mi procedure.

Name	Description
bwt	birth weight (g)
female	sex of child (1=female, 0=male)
plurality	number of children at delivery (1=single, 2=twins, : : :, 5=quintuplets or higher)
gestation	number of weeks (17; 18; : : : ; 47)
birth_order	birth order (1=1st child, 2=2nd child, : : :, 8=8th or higher child)
mother_age	mother's age (year)
mother_white	mother's race (1=white, 0=otherwise)
mother_college	mother's education (1=at least college degree, 0=no college degree)
mother_single	mother's marital status (1=single, 0=married)
mother_bmi	mother's prepregnancy bmi (kg/m2)
mother_smoking	mother's smoking status (1=yes, 0=no)
wic	WIC receipt (1=yes, 0=no)
prenatal	number of prenatal care visits (0; 1; 2; : : : ; 49)
resident	resident status (1=U.S. resident, 0=foreign resident)
father_age	father's age (year)
father_white	father's race (1=white, 0=otherwise)
father_college	father's education (1=at least college degree, 0=no college degree)

Question 1

Part A

Using an O'Brien-Fleming sequential plan for detecting early evidence of superior efficacy, write decision rules that can be included in the data monitoring committee interim analysis guidelines (also known as the data monitoring committee charter). The plan is to conduct 5 interim analyses plus 1 final look (6 total looks), and assume a two-sided 5% significance level.

The O'Brien-Fleming approach is a common group sequential approach. In this case, we will conduct 6 analyses (5 interim and 1 final look). We will use a two-sided 5% significance level (overall $alpha = 0.05$). We will set a $K = 5$. This would set a series of boundaries of for each iteration:

1. $Z_1 = 4.56$
2. $Z_2 = 3.23$
3. $Z_3 = 2.63$
4. $Z_4 = 2.28$
5. $Z_5 = 2.04$

We would stop analysis if the interim analysis value crossed the boundaries established by the O'brien-Fleming approach.

Part B

Using a Pocock sequential plan, write the corresponding decision rules.

If a Pocock sequential plan was used, then a similar interim significance level would be used to maintain an overall $\alpha = 0.05$ approach. As there are 5 interim analyses, we would use $C_i = 2.4$. If the study crosses this, then we can terminate the study and reject the H_0.

Part C

Compare the O'Brien-Fleming and Pocock stopping rules.

The Pocock stopping rules include that hte final analysis is conducted at a much smaller alpha than an $\alpha = 0.05$. The O'Brien-Fleming approach allows for the final analysis to have a higher α, but requires more conservatism (higher critical values) in the earlier interim analyses.

Question 2

Part A

When performing a complete case analysis, how many cases will be used? How many cases will be ignored?

Out of this data set, the complete cases are 19195. There are overall 19984 cases, thus the cases that would be ignored are 789

Part B

Write a multiple linear regression model to find an estimate of the difference in birth weight between smoking mothers and non-smoking mothers, holding plurality, mother age, gestation, and mother white fixed. Provide a 95% confidence interval.

The overall model is:

$$BirthWeight = \beta_0 + \beta_1 MotherSmoking + \beta_2 Plurality + \beta_3 MotherAge + \beta_4 Gestation + \beta_5 MotherWhite$$

Multiple Linear Regression for Birthweight

	Dependent variable:
	bwt
mother_smoking	-122.810^{***}
	(-143.820, -101.790)
plurality	-546.330^{***}
	(-583.170, -509.500)
mother_age	7.414^{***}
	(6.294, 8.535)
gestation	121.410^{***}
	(118.610, 124.220)
mother_white	129.890^{***}
	(114.170, 145.610)
Constant	$-1{,}147.300^{***}$
	($-1{,}274.400$, $-1{,}020.300$)
Observations	19,195
R^2	0.354
Adjusted R^2	0.354
Residual Std. Error	473.440 (df = 19189)
F Statistic	$2{,}103.300^{***}$ (df = 5; 19189)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Question 3

Part A

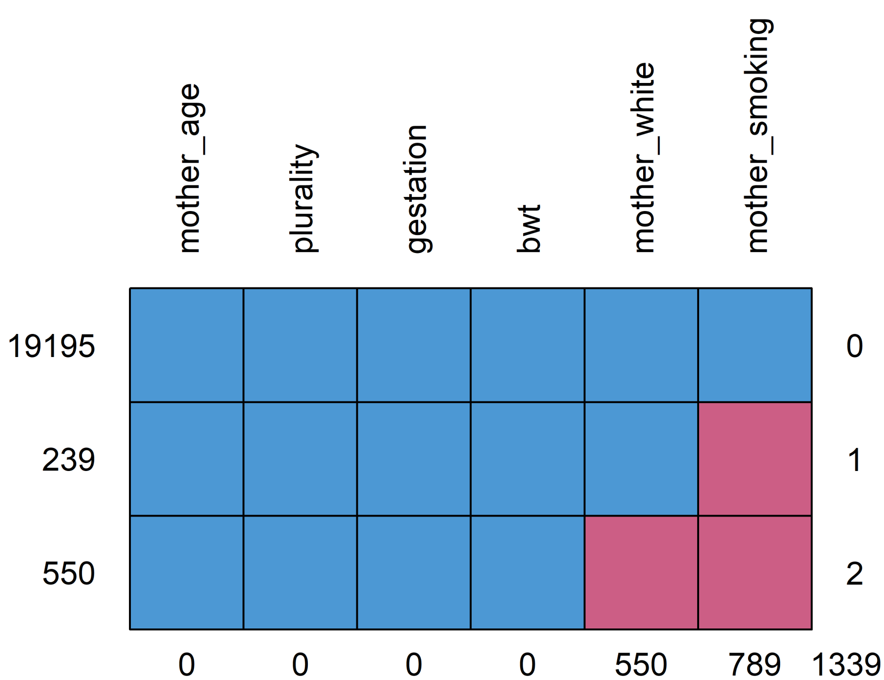
Summarize each variable's missingness by reporting the number of missing values for the variable and the percentage missing.

Missingess table

	[ALL] N=19984	N
bwt	0 (0.00%)	19984
mother_smoking	789 (3.95%)	19984
plurality	0 (0.00%)	19984
mother_age	0 (0.00%)	19984
gestation	0 (0.00%)	19984
mother_white	550 (2.75%)	19984

Part B

Is the pattern of missingness monotone or arbitrary? What is the most prevalent pattern of missingness in this dataset?



Missingness Pattern

	mother_age	plurality	gestation	bwt	mother_white	mother_smoking	
19195	1	1	1	1	1	1	0
239	1	1	1	1	1	0	1
550	1	1	1	1	0	0	2
	0	0	0	0	550	789	1339

In this data set, the missingness is monotonic. When **mother_white** is missing, **mother_smoking** is missing as well, for a total of 550 observations. If **mother_smoking** is missing only, there are 239 observations.

Question 4

Single vs multiple imputation. Why is multiple imputation preferred over single imputation?

Single imputation leads to replacing a missing value with some other value (the mean, etc). Multiple imputation fills the missing value with a range of plausible values (allowing the inclusion of uncertainty). The number of multiple imputations (e.g. 3 sets of data, 15 sets, etc) are then anlayzed together giving a range of results, which are more likely to contain the true sample/population parameters.

Question 5

Missing data mechanisms. Multiple imputation assumes MAR. How is MAR different from MCAR? How is MAR different from NMAR?

MAR is missing at random, while MCAR is missing completely at random, and NMAR is not missing at random. MAR suggests that the probability of the data missing is unrelated to its value, however the missing data may be related to ther varuables. MCAR means that missingness is unrelated to the values of any other variables (present or missing), and is usually an overly strong assumption. MAR is different from NMAR, suggests that the value of hte unobserved varaible itself predicts its own missingness, thus cannot be ignored.

Question 6

Perform multiple imputation to create $m = 6$ complete data sets. Be sure to include mother white and mother smoking in the CLASS statement, and use the LOGISTIC function to impute mother white and mother smoking after the MONOTONE statement. Use the following order of the variables: bwt, plurality, mother age, gestation, mother white, mother smoking. Recall that SAS will follow the variable order that you specify. Copy and paste the Missing Data Patterns table generated by SAS.

Imputation Patterns

	mother_age	plurality	gestation	bwt	mother_white	mother_smoking
mother_age	0	1	1	1	1	1
plurality	1	0	1	1	1	1
gestation	1	1	0	1	1	1
bwt	1	1	1	0	1	1
mother_white	1	1	1	1	0	1
mother_smoking	1	1	1	1	1	0

Question 7

Analysis on imputed data sets. Build a linear regression model using bwt as outcome and mother smoking as primary exposure, using plurality, mother age, gestation, mother white, as control variables. Because there are 6 imputed data sets, you will have 6 sets of

s. Combine the results of the 6 imputations.

Part A

Report an estimate of the difference in birth weight between smoking mothers and non-smoking mothers, holding plurality, mother age, gestation, and mother white fixed? Provide a 95% confidence interval.

Part B

Conclusion. Compare the results obtained using multiple imputation and those obtained using complete case analysis. Describe your observations.

Here is without imputation:

Model with Missing

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1147.3446	64.80639	-17.704	0	-1274.3708	-1020.3184
mother_smoking	-122.8061	10.72271	-11.453	0	-143.8236	-101.7887
plurality	-546.3321	18.79481	-29.068	0	-583.1716	-509.4926
mother_age	7.4144	0.57179	12.967	0	6.2936	8.5351
gestation	121.4140	1.42919	84.953	0	118.6127	124.2154
mother_white	129.8909	8.02192	16.192	0	114.1672	145.6146

Here is with multiple imputation:

Combined Regressions for the Multiple Imputations (x6)

term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-1132.6366	63.34798	-17.880	19922.6	0	-1256.8040	-1008.4693
mother_smoking	-122.8184	10.58943	-11.598	10524.3	0	-143.5757	-102.0611
plurality	-550.5448	18.18444	-30.276	19958.0	0	-586.1878	-514.9018
mother_age	7.2614	0.56028	12.960	19745.8	0	6.1632	8.3596
gestation	121.2308	1.39918	86.644	19941.9	0	118.4882	123.9733
mother_white	130.8674	7.99346	16.372	3557.8	0	115.1951	146.5396

The estimates with missing and imputed data are very similar. The effect of **mother_smoking** and **mother_white** on **bwt** remains with almost completely overlapping confidence intervals. With imputation, the confidence intervals may be slightly narrowed (as the estimates were more powered).