94-827: SAS for Public Policy/Policy Analysis

Final 2022

Due 2022-03-04

121 points

The final exam has 4 parts. Part 1 requires you to read output and interpret the output. In Part 2, you must answer questions about existing code and answer questions about what the program does. Answers to Part 1 questions and Part 2 questions should be answered in comments.

Parts 3 and 4 are related. You are given a SAS program in Part 3 and you must run the code, and add descriptive comments. The output datasets of the program in Part 3 will be needed to complete your analysis in Part 4. The data and programs needed for the final can be found on Canvas and in the course folder under SASOnDemand. You may use lecture notes, SAS programs saved online/Canvas, your books, and the internet. Your work must be your own.

For Part 4, data sources will need to be prepared before getting to perform exploratory data analysis, run models and interpret results. All parts in RED should be answered and included in your program. Part 4 also includes graphical output. Export your final graphs to a PDF (with titles). Each student should turn in one PDF (from Part 4), one SAS program with the answers for Parts 1 and 2, and one SAS program for Parts 3 and 4 with the code and answers.

Part 1: Reading Output (30 points)

This SAS program called Parts_1and2.sas can be found on Canvas. Save the program to a personal location and answer the questions in comments. The datasets are in the FINAL library. You may run the program if you find it helpful. You may run other procedures to help you gain more information. If you do, include that code with your answers.

```
proc logistic data=Final.LoanApp;
        class loanpurp mortapp/param=reference;
        model loanapproved = loanpurp mortapp creditscore price;
run;
```

Model Information		
Data Set	FINAL.LOAN APP	
Response Variable	LoanApproved	Loan Approved (0=no, 1=yes)
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	4999
Number of Observations Used	4999

Response Profile

Ordered Value	LoanApproved	Total Frequency
1	0	1880
2	1	3119

Probability modeled is LoanApproved=0.

Class Level Information					
Class	Value	Design Variables			
LoanPurp	1				
MortApp	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6621.776	5942.847
SC	6628.293	5988.466
-2 Log L	6619.776	5928.847

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	690.9290	6	<.0001
Score	674.2830	6	<.0001
Wald	578.6876	6	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LoanPurp	0	.	.
MortApp	4	0.5195	0.9716
CreditScore	1	1.0123	0.3144
Price	1	577.7748	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1944	0.3683	35.5042	<.0001
MortApp	1	1	0.0292	0.1450	0.0404	0.8406
MortApp	2	1	0.0234	0.1719	0.0186	0.8915
MortApp	3	1	0.0661	0.1961	0.1137	0.7359
MortApp	4	1	0.1230	0.2018	0.3714	0.5422
CreditScore		1	0.000470	0.000467	1.0123	0.3144
Price		1	0.00205	0.000085	577.7748	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MortApp 1 vs 5	1.030	0.775	1.368
MortApp 2 vs 5	1.024	0.731	1.434
MortApp 3 vs 5	1.068	0.727	1.569
MortApp 4 vs 5	1.131	0.761	1.680
CreditScore	1.000	1.000	1.001
Price	1.002	1.002	1.002

```
proc ttest data=Final.Study_GP;
        class section;
        var GPA;
run;
```

Section	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
01		58	3.3014	0.3941	0.0517	2.4200	3.9400
02		64	3.1013	0.4494	0.0562	1.9300	3.9100
Diff (1-2)	**Pooled**		0.2001	0.4240	0.0769		
Diff (1-2)	**Satterthwaite**		0.2001		0.0764		

Section	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
01		3.3014	3.1978	3.4050	0.3941	0.3332	0.4825
02		3.1013	2.9890	3.2135	0.4494	0.3828	0.5443
Diff (1-2)	**Pooled**	0.2001	0.0479	0.3523	0.4240	0.3765	0.4854
Diff (1-2)	**Satterthwaite**	0.2001	0.0489	0.3514			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	120	2.60	0.0104
Satterthwaite	Unequal	119.88	2.62	0.0099

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	63	57	1.30	0.3153

```
proc freq data=Final.birthwgt;
	table Drinking * lowbirthwgt;
	table AgeGroup * lowbirthwgt;
run;
```

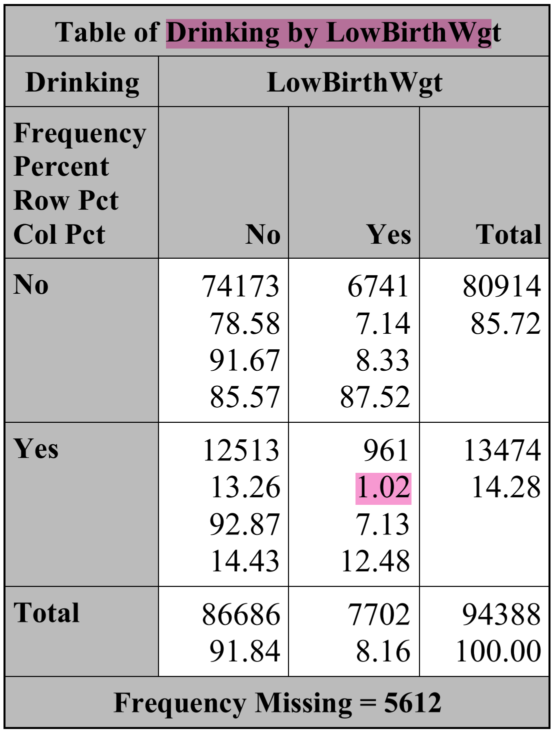


Table of Drinking by LowBirthWgt

Drinking	LowBirthWgt		
Frequency Percent Row Pct Col Pct	No	Yes	Total
No	74173 78.58 91.67 85.57	6741 7.14 8.33 87.52	80914 85.72
Yes	12513 13.26 92.87 14.43	961 1.02 7.13 12.48	13474 14.28
Total	86686 91.84	7702 8.16	94388 100.00

Frequency Missing = 5612

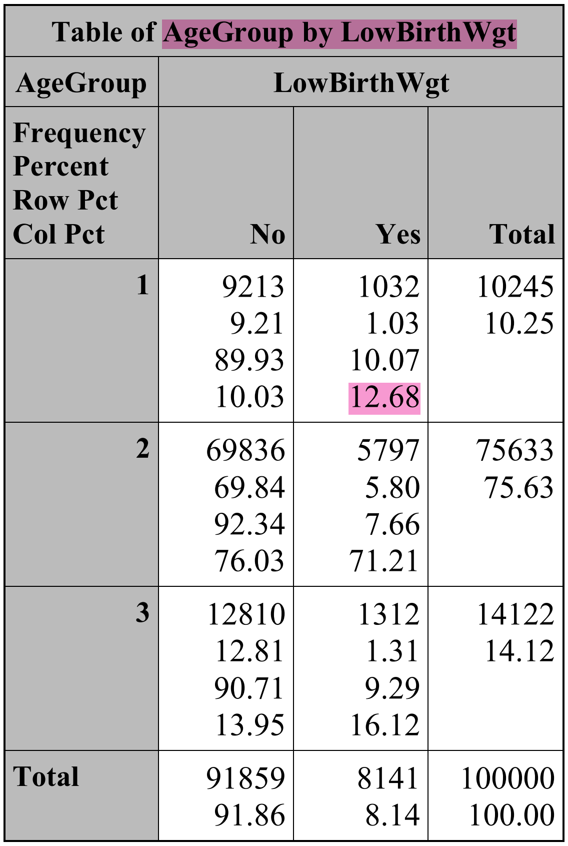


Table of AgeGroup by LowBirthWgt

AgeGroup	LowBirthWgt		
Frequency Percent Row Pct Col Pct	No	Yes	Total
1	9213 9.21 89.93 10.03	1032 1.03 10.07 12.68	10245 10.25
2	69836 69.84 92.34 76.03	5797 5.80 7.66 71.21	75633 75.63
3	12810 12.81 90.71 13.95	1312 1.31 9.29 16.12	14122 14.12
Total	91859 91.86	8141 8.14	100000 100.00

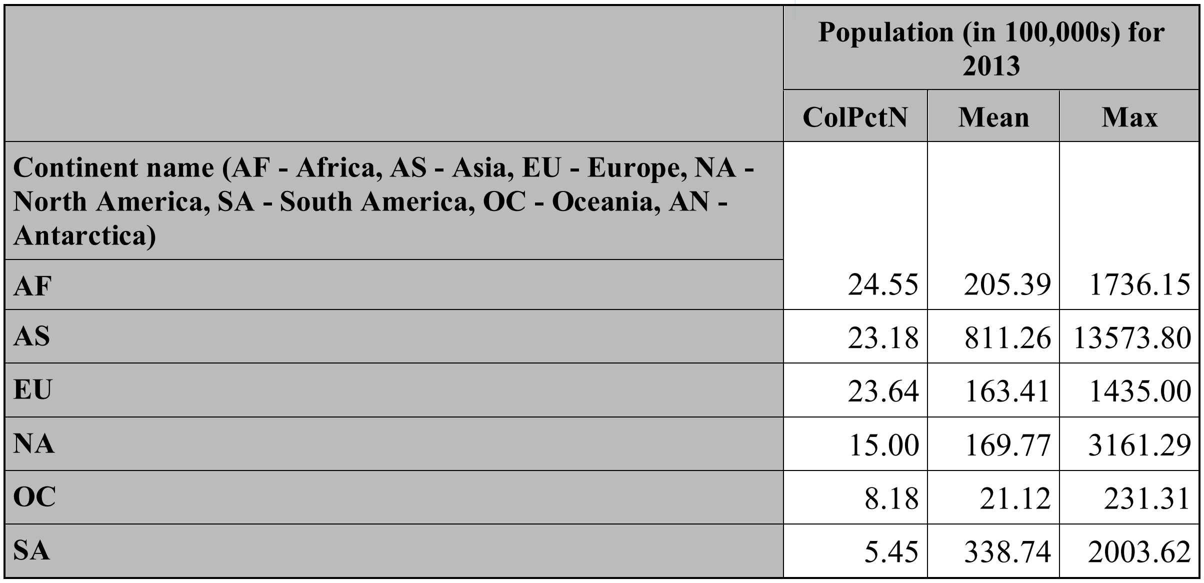
```
proc corr data=Final.Vite;
        var plaque sbp ldl hdl trig;
run;
```

5 Variables:	Plaque SBP LDL HDL Trig

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
Plaque	1500	0.63290	0.17252	949.35040	0.22090	1.08080	Plaque measurement (mm)
SBP	1500	141.87533	27.36373	212813	65.00000	234.00000	Systolic blood pressure (mm/Mg)
LDL	1500	135.52800	14.98582	203292	83.00000	185.00000	LDL cholesterol (mg/DL)
HDL	1500	45.86533	6.82776	68798	22.00000	71.00000	HDL cholesterol (mg/DL)
Trig	1500	173.56067	87.72554	260341	25.00000	503.00000	triglycerides mg/dL

Pearson Correlation Coefficients, N = 1500 Prob > |r| under H0: Rho=0					
	Plaque	**SBP**	**LDL**	**HDL**	**Trig**
Plaque **Plaque measurement (mm)**	1.00000	-0.01398 0.5885	0.00029 0.9911	-0.13821 <.0001	0.03225 0.2119
SBP **Systolic blood pressure (mm/Mg)**	-0.01398 0.5885	1.00000	0.00766 0.7669	-0.00086 0.9733	-0.03266 0.2062
LDL **LDL cholesterol (mg/DL)**	0.00029 0.9911	0.00766 0.7669	1.00000	-0.01074 0.6777	0.03352 0.1945
HDL **HDL cholesterol (mg/DL)**	-0.13821 <.0001	-0.00086 0.9733	-0.01074 0.6777	1.00000	0.02471 0.3388
Trig **triglycerides mg/dL**	0.03225 0.2119	-0.03266 0.2062	0.03352 0.1945	0.02471 0.3388	1.00000

```
proc tabulate data=Final.Population;
	class Continent;
	variable y1;
	table Continent, Y1*(ColPctN Mean Max);
run;
```



	Population (in 100,000s) for 2013		
	ColPctN	Mean	Max
Continent name (AF - Africa, AS - Asia, EU - Europe, NA - North America, SA - South America, OC - Oceania, AN - Antarctica)			
AF	24.55	205.39	1736.15
AS	23.18	811.26	13573.80
EU	23.64	163.41	1435.00
NA	15.00	169.77	3161.29
OC	8.18	21.12	231.31
SA	5.45	338.74	2003.62

Part 2:Analyzing Code (12 points)
In Part 2, answer the questions (1 to 8) regarding data steps and procedures using the dataset FINAL.CARS. Answer the questions using comments and submit them to Canvas.

Parts 3 and 4: Data Preparation, Original Program, and Policy Analysis (79 points)

Effects of marijuana dispensaries on crime and property transaction values

In 2010, Colorado legalized medical marijuana centers, after legalizing medical marijuana usage in 2000. Later in 2012, Colorado became one of the first states to end the ban on recreational marijuana. There have been several studies that look at the impacts of marijuana legalization, and specifically of the dispensaries. These studies are often geospatial in nature, analyzing incidents or conditions near the dispensaries. We will look at one study examining the location of dispensaries and crime incidents in Denver, Colorado for the final exam:

> *Marijuana Dispensaries and Neighborhood Crime and Disorder in Denver, Colorado.* Lorine A. Hughes, Lonnie M. Schaible, and Katherine Jimmerson. Justice Quarterly (2019)

The paper is provided for your reference and may help with understanding why and how some steps should be completed. The final will not directly replicate the steps done in the paper due to time constraints and complexity. Some of the tables will be referenced, but is not necessary to read the entire paper.

Crime and Disorder (2019) concludes that "except for murder, the presence of at least one medical marijuana dispensary was associated with statistically significantly increased neighborhood crime and disorder." Using the methods you've learned, you will analyze some of the same data sources and indicate whether you agree with the conclusions of the model paper.

Answers to questions and descriptions of the process can be typed into your program as comments. Six (6) points of your score reflects good coding practices, such as descriptive comments, usage of titles for procedures, and checks on new data or newly derived variables. (6 points)

During this portion of the final you should use, at least once (in no particular order):

- PROC SORT
- IF…THEN
- New variable creation
- IN=
- SAS functions
- ODS Output
- PROC CONTENTS
- Merging
- PROC MEANS or TABULATE
- PROC FREQ
- Any other procedure explicitly mentioned

Description of datasets

1. Unfortunately, the data for this analysis are in several datasets and in different formats. This is common when working with data, particularly administrative datasets that were not created for the sake of analysis like survey data. Below, each file needed to complete the analysis is listed.
    a. *Tract_medical_disp_totals.sas7bdat*. This is a SAS dataset that has the number of medical marijuana dispensaries in each census tract of Denver for the years 2010 to 2019. For example, a tract may have had 2 dispensaries in 2016, and the same tract may have had 4 dispensaries in 2019.
    b. *Tract_retail_disp_totals.sas7bdat*. This is a SAS dataset that has the number of retail marijuana dispensaries in each census tract of Denver for each year from 2010 to 2019.
    c. *Years.sas7bdat.* This is a SAS dataset that lists each year from 2010 to 2019.
    d. *Tract_nhood.sas7bdat.* This SAS dataset links each census tract to a neighborhood in Denver. Each census tract is fully contained in one neighborhood, but some neighborhoods contain multiple census tracts.
    e. *Crime.csv.* The crime file contains incident level information on each crime in Denver from 2015 through part of 2020. The neighborhood in which the crime occurred is listed. This file will be imported for you by running Part3.sas.
    f. *ACS2018_DenverTracts.sas7bdat.* This file contains selected demographic information for all of the tracts in Denver based on the ACS 5 year tract estimates from 2014 to 2018. The list of demographics was based on the *Crime and Disorder* paper.

Crime Data Preparation

2. Run Part3.sas and respond to the prompts (A – Q) in your comments. (23 points)

3. Merge the datasets containing retail dispensary totals, the medical dispensary totals, and TRACT_NHOOD_YEAR together by the fips code and year. If a tract does not have a record for a given year for either of the dispensary files, assume that the tract had 0 dispensaries of that type for that year. (3 points)

4. The analysis is *Crime and Disorder* is at the neighborhood level. Use appropriate procedures to calculate the total number of each dispensary type in each neighborhood in each year, using the file created in #3. Your output dataset (call it NEIGHBORHOOD_DISPENS) from this step should have one record per neighborhood – year combination. Next, calculate 3 binary variables, where 1 indicates there is at least 1 dispensary of that type in the neighborhood: (4 points)
    a. Has_retail_dispensary
    b. Has_medical_dispensary
    c. Has_dispensary (at least one of any type)

5. Merge CRIME03 and CRIME_MONTH_NEIGHBORHOOD together by neighborhood, year, month, and crime category. If a neighborhood does not have a record for a crime category in a given month, assume that the neighborhood had 0 crimes of that category for that month. Keep only the crimes listed in Table 1 of *Crime and Disorder* and keep only records from 2015 to 2019 in your dataset. (4 points)
    a. **Was this a 1 to 1, 1 to Many, or Many to Many merge?**

6. Merge the dataset you created in #5 with NEIGHBORHOOD_DISPENS using neighborhood and year. Ensure the resulting files only has records between 2015 and 2019. We are attaching the number of dispensaries in a given year for a neighborhood, to that neighborhoods monthly crime totals. (3 points)

7. Merge the dataset created in #6 with NHOOD_3 using neighborhood. The period over which the ACS data were collected, roughly aligns with the period in which crime data were collected. Using the crime totals and neighborhood population, compute crime rate (# of crimes per 100,000 people). This is your CRIME_ANALYTIC_DATASET. (3 points)

Analysis

8. Using appropriate procedures and your analytic dataset produce tables that show descriptive statistics similar to Table 1 of *Crime and Disorder.* (3 points)

9. Using appropriate procedures, create two charts of your choosing and output them to a PDF. Using footnotes, write 1 to 3 sentences describing what the charts show. Be sure to submit your PDF on Canvas! (5 points)

10. Use appropriate procedures to determine if the crime rate you calculated for given crime types are (approximately) normally distributed. EACH CRIME TYPE SHOULD BE CONSIDERED SEPARATELY. (4 points)

11. Use appropriate procedures to determine if any of the ACS neighborhood demographics are associated with crime rates or monthly totals for Public Disorder. (4 points)
    a. **List any significant correlations in your comments.**

12. Macro creation. Create and invoke a macro that accomplishes the following. It does not have to have a %DO loop, but if that is your preference, that is fine. (6 points)
    a. Use PROC GLM to run a linear regressions for EACH crime category, where the crime rate is the dependent variable in one set, and monthly totals is the dependent variable in the other set. The independent variables should include the number of each type of dispensary in the neighborhood. Crime has seasonal components (higher in some months while lower in others). Include in your regression dummy variables for months and years (HINT: These are categorical in

this case). Finally include the neighborhood demographic variables you believe make sense. (6 points)

b. AT THE END OF YOUR MODEL STATEMENT, MAKE SURE YOU INCLUDE "/ solution" before the semi-colon. If you do not, SAS will not produce parameter estimates.

13. Discuss significant findings related to the demographics and your variables of interest. There will be a lot of regressions (16) and lots of regression coefficients. Some patterns should emerge. Don't try to discuss everything individually. If you want to focus on just one, choose Public Disorder. This discussion does not have to be long (2 to 4 paragraphs), and if you prefer to write it in Word, that is fine. Some guiding questions: (11 points)

    a. Based on the R-squared, generally speaking which set of models explains more of the variation: crime rates or monthly totals?
    b. Which crime types seem positively related to the number of retail dispensaries? Which are negatively related? Which are unrelated?
    c. Choose one of the demographic characteristics. Discuss whether it is significantly related to the any or most of the crime rates.
    d. In some cases, are there differences between the effects of medical versus retail dispensaries?
    e. Do you agree with the conclusion of the paper?