VIRTUAL

SAS® GLOBAL FORUM 2021

#SASGF

Back to Basics: Running an Analysis from Data to Refinement in SAS

Deanna Schreiber-Gregory, Juxdapoze, LLC

Deanna is a Lead Research Statistician and Data Manager on contract through the Henry M Jackson Foundation for the Advancement of Military Medicine to the Department of Defense in Bethesda, MD. She is also an Independent Consultant for Statistics, Research Methods, and Data Management in the private sector through Juxdapoze, LLC. Deanna has an MS in Health and Life Science Analytics, a BS in Statistics, and a BS in Psychology. Deanna has presented as a contributed and invited speaker at over 50 local, regional, national, and global SAS user group conferences since 2011.

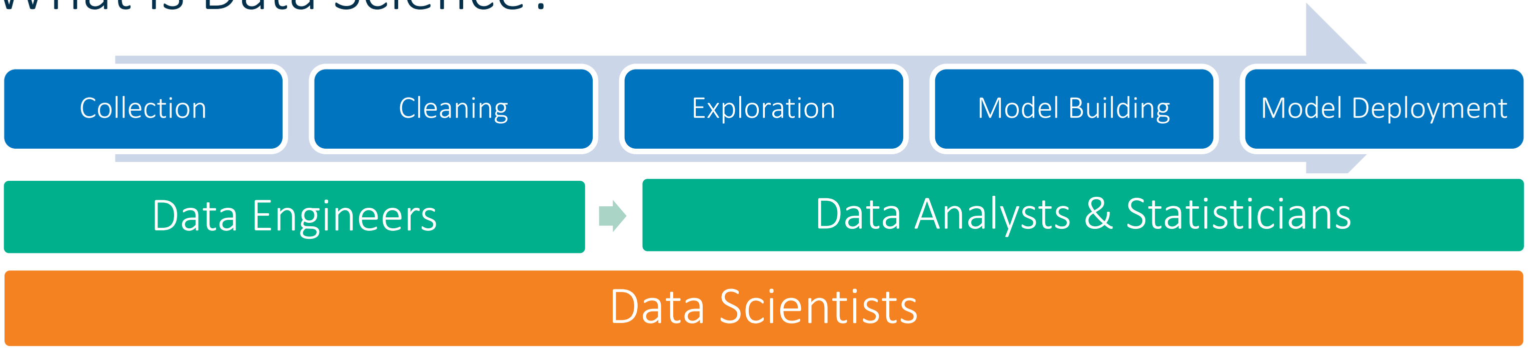
Data Science

The [Cyber]Space Race

Introduction

The Analytic Process

- What is Data Science?

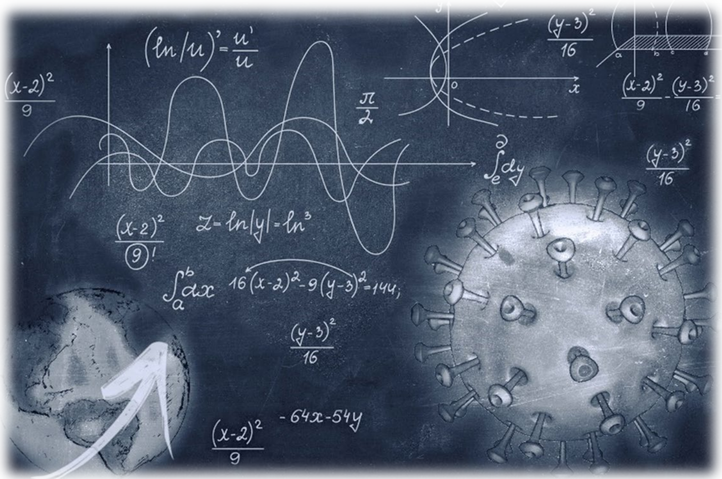


- The COVID Puzzle
  - Limited timeframe & Real-time information
  - Inconsistent reporting, sources, and standards
  - Little to no prior knowledge to work off of
  - Fragile/inconsistent reporting, sources, and standards

Introduction

Our Subject: COVID Data

- Lots of Sources
  - CDC COVID Tracker
  - HealthData.gov
  - WHO Global Table
  - Kaggle COVID 19 Challenge
  - NIH Open-Access COVID 19
  - Individual state-based repositories
  - John’s Hopkins COVID 19 Secondary Analysis
  - Open ICPSR COVID-19 Data Repository
  - MIDAS Online Portal for COVID-19 Modeling Research
  - Google Cloud – Big Query – COVID 19 public datasets



Outline

- Choosing and Importing Data
- Data Exploration
- Data Driven Modeling
- Matching Your Question to a Model
- Evaluate Your Model
- Conclusion

Choosing and Importing Data

CDC COVID Tracker

Choosing and Importing Data

Key Considerations

- Choosing Data
  - Primary Analyses vs Secondary Analyses
  - Experimental vs Observational Research
  - Does the data ask/answer the right questions
  - What is the data structure – survey, clinical, observational/mined
  - Sample size / generalizability
- Importing Data
  - Data comes in a variety of formats – Excel, SPSS, txt, etc
  - Some data may need to be merged

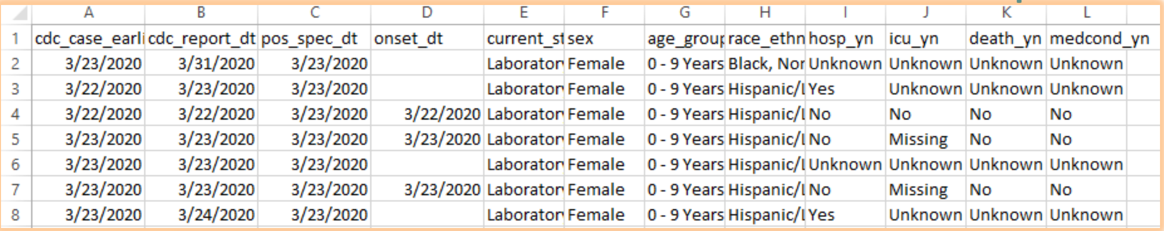
Choosing and Importing Data

SAS Procedures

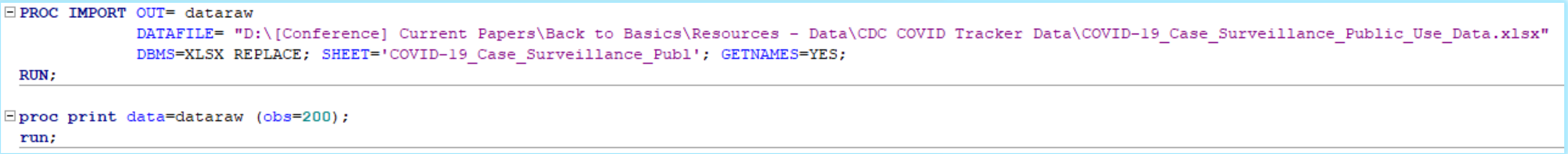
- Getting Data Into SAS
  - Can Enter It Directly
  - Proc Import
- Consider Data Storage
  - Can store data as a SAS dataset (permanent or temporary)
  - Can export data in your desired format (Proc Export)

Choosing and Importing Data

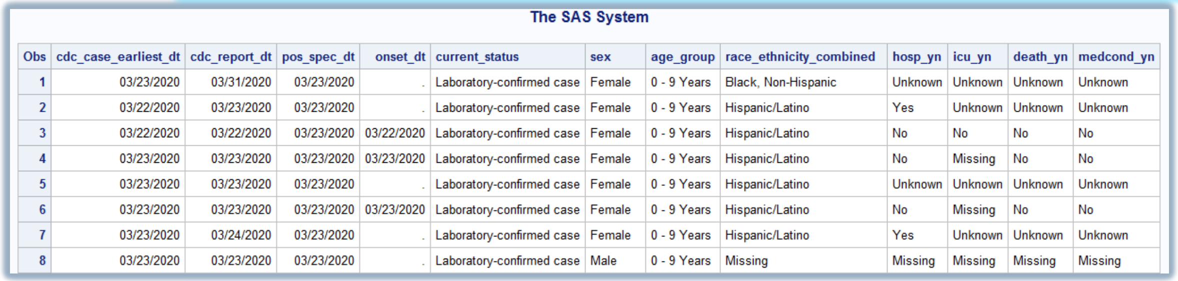
COVID Example



	A	B	C	D	E	F	G	H	I	J	K	L
1	cdc_case_earli	cdc_report_dt	pos_spec_dt	onset_dt	current_st	sex	age_grou	race_ethn	hosp_yn	icu_yn	death_yn	medcond_yn
2	3/23/2020	3/31/2020	3/23/2020		Laborator	Female	0 - 9 Years	Black, Nor	Unknown	Unknown	Unknown	Unknown
3	3/22/2020	3/23/2020	3/23/2020		Laborator	Female	0 - 9 Years	Hispanic/l	Yes	Unknown	Unknown	Unknown
4	3/22/2020	3/22/2020	3/23/2020	3/22/2020	Laborator	Female	0 - 9 Years	Hispanic/l	No	No	No	No
5	3/23/2020	3/23/2020	3/23/2020	3/23/2020	Laborator	Female	0 - 9 Years	Hispanic/l	No	Missing	No	No
6	3/23/2020	3/23/2020	3/23/2020		Laborator	Female	0 - 9 Years	Hispanic/l	Unknown	Unknown	Unknown	Unknown
7	3/23/2020	3/23/2020	3/23/2020	3/23/2020	Laborator	Female	0 - 9 Years	Hispanic/l	No	Missing	No	No
8	3/23/2020	3/24/2020	3/23/2020		Laborator	Female	0 - 9 Years	Hispanic/l	Yes	Unknown	Unknown	Unknown

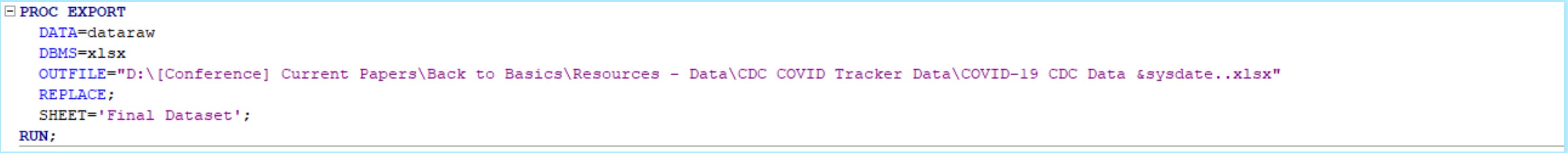


```
PROC IMPORT OUT= dataraw
            DATAFILE= "D:\[Conference] Current Papers\Back to Basics\Resources - Data\CDC COVID Tracker Data\COVID-19_Case_Surveillance_Public_Use_Data.xlsx"
            DBMS=XLSX REPLACE; SHEET='COVID-19_Case_Surveillance_Publ'; GETNAMES=YES;
RUN;

proc print data=dataraw (obs=200);
run;
```



The SAS System

Obs	cdc_case_earliest_dt	cdc_report_dt	pos_spec_dt	onset_dt	current_status	sex	age_group	race_ethnicity_combined	hosp_yn	icu_yn	death_yn	medcond_yn
1	03/23/2020	03/31/2020	03/23/2020	.	Laboratory-confirmed case	Female	0 - 9 Years	Black, Non-Hispanic	Unknown	Unknown	Unknown	Unknown
2	03/22/2020	03/23/2020	03/23/2020	.	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	Yes	Unknown	Unknown	Unknown
3	03/22/2020	03/22/2020	03/23/2020	03/22/2020	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	No	No	No	No
4	03/23/2020	03/23/2020	03/23/2020	03/23/2020	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	No	Missing	No	No
5	03/23/2020	03/23/2020	03/23/2020	.	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	Unknown	Unknown	Unknown	Unknown
6	03/23/2020	03/23/2020	03/23/2020	03/23/2020	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	No	Missing	No	No
7	03/23/2020	03/24/2020	03/23/2020	.	Laboratory-confirmed case	Female	0 - 9 Years	Hispanic/Latino	Yes	Unknown	Unknown	Unknown
8	03/23/2020	03/23/2020	03/23/2020	.	Laboratory-confirmed case	Male	0 - 9 Years	Missing	Missing	Missing	Missing	Missing



```
PROC EXPORT
   DATA=dataraw
   DBMS=xlsx
   OUTFILE="D:\[Conference] Current Papers\Back to Basics\Resources - Data\CDC COVID Tracker Data\COVID-19 CDC Data &sysdate..xlsx"
   REPLACE;
   SHEET='Final Dataset';
RUN;
```

Choosing and Importing Data

Best Practices

BEST PRACTICE

- Choosing Data
  - Pay attention to sources
  - Understand that data has limitations
- Importing Data
  - Make sure there is no excess information in the datasheet
  - Variables should have names that make sense
  - Use labels instead of long names
  - Maintain an untouched original dataset without adjustments
  - Document your adjustments
  - Always review your log

Data Exploration

Health Data.gov Hospital Data

Data Exploration

Key Considerations

- Consider Data Types
  - Numeric
    - Continuous vs Discrete
    - Interval vs Ratio
  - Categorical
    - Binary/Dichotomous vs Multi-level
    - Nominal vs Ordinal
    - Dummy
- Data Cleaning
  - IMPORTANT
  - Natural part of this step

Numeric	Interval	Ratio
Discrete	Calendar Years 100 BC, 100 AD, 2019 AD	# Children 0, 1, 2, 3, 4
Continuous	Temperature -10.1˚, 0˚, 10.9˚, 20˚	Height 0.2ft, 1.2ft, 2.2ft

Categorical	Ordinal	Nominal
Binary / Dichotomous	Pass Fail	Male Female
Multi-level	Honors Pass Marginal Pass Fail	Male Female Trans-Male Trans-Female

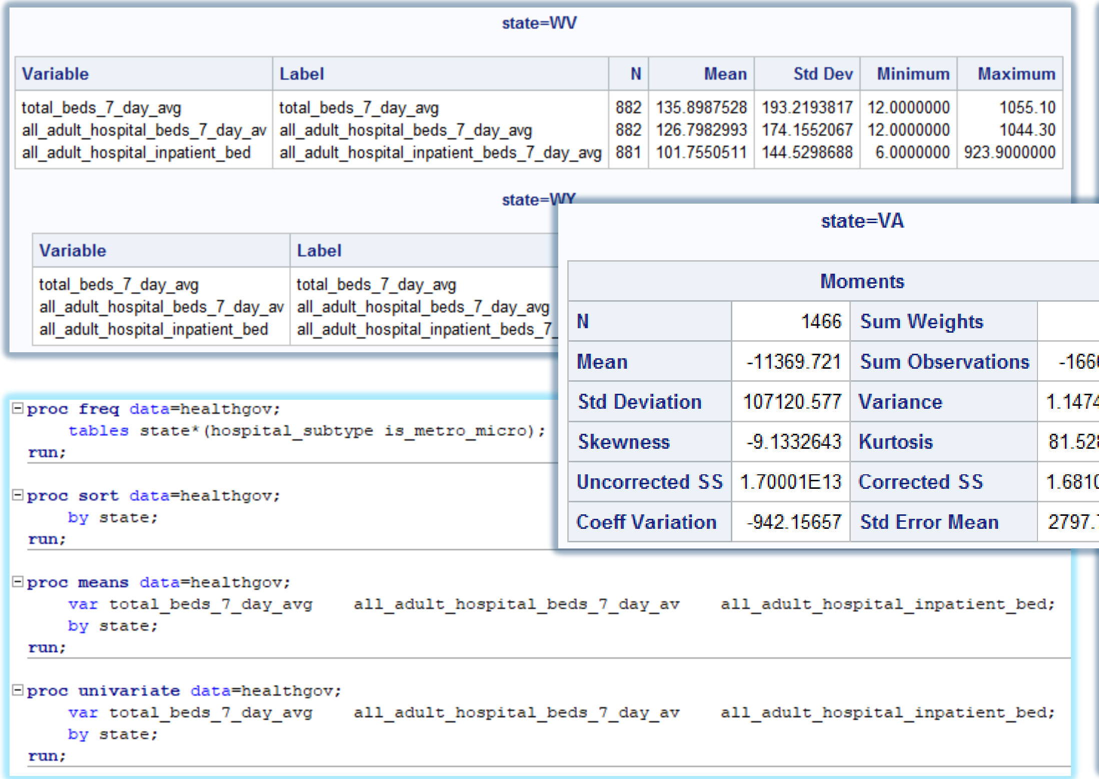
Data Exploration

SAS Procedures

- SAS Procedures
  - Proc Freq
  - Proc Means
  - Proc Univariate
  - Proc Corr
- The Data Step
  - Implement adjustments to the data
- Additional Helpful Procedures
  - Proc Contents
  - Proc Sort
  - Proc SQL
  - Proc Print
  - SAS Macros

Data Exploration

COVID Example



state=WV

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
total_beds_7_day_avg	total_beds_7_day_avg	882	135.8987528	193.2193817	12.0000000	1055.10
all_adult_hospital_beds_7_day_av	all_adult_hospital_beds_7_day_avg	882	126.7982993	174.1552067	12.0000000	1044.30
all_adult_hospital_inpatient_bed	all_adult_hospital_inpatient_beds_7_day_avg	881	101.7550511	144.5298688	6.0000000	923.9000000

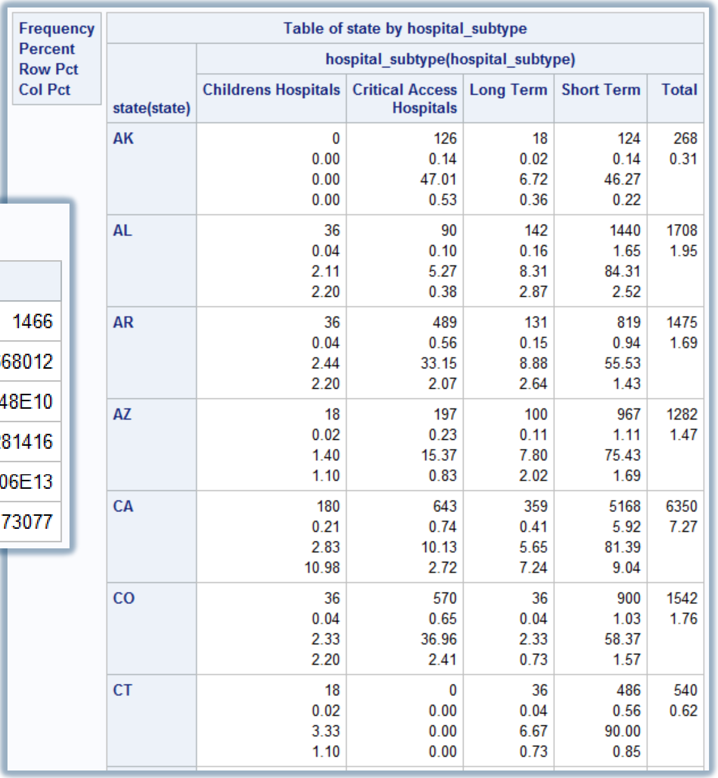
state=WY

Variable	Label
total_beds_7_day_avg	total_beds_7_day_avg
all_adult_hospital_beds_7_day_av	all_adult_hospital_beds_7_day_avg
all_adult_hospital_inpatient_bed	all_adult_hospital_inpatient_beds_7_

```
proc freq data=healthgov;
    tables state*(hospital_subtype is_metro_micro);
run;

proc sort data=healthgov;
    by state;
run;

proc means data=healthgov;
    var total_beds_7_day_avg    all_adult_hospital_beds_7_day_av    all_adult_hospital_inpatient_bed;
    by state;
run;

proc univariate data=healthgov;
    var total_beds_7_day_avg    all_adult_hospital_beds_7_day_av    all_adult_hospital_inpatient_bed;
    by state;
run;
```

state=VA

Moments			
N	1466	Sum Weights	1466
Mean	-11369.721	Sum Observations	-16668012
Std Deviation	107120.577	Variance	1.14748E10
Skewness	-9.1332643	Kurtosis	81.5281416
Uncorrected SS	1.70001E13	Corrected SS	1.68106E13
Coeff Variation	-942.15657	Std Error Mean	2797.73077



Frequency
Percent
Row Pct
Col Pct

Table of state by hospital_subtype

	hospital_subtype(hospital_subtype)				
state(state)	Childrens Hospitals	Critical Access Hospitals	Long Term	Short Term	Total
AK	0 0.00 0.00 0.00	126 0.14 47.01 0.53	18 0.02 6.72 0.36	124 0.14 46.27 0.22	268 0.31
AL	36 0.04 2.11 2.20	90 0.10 5.27 0.38	142 0.16 8.31 2.87	1440 1.65 84.31 2.52	1708 1.95
AR	36 0.04 2.44 2.20	489 0.56 33.15 2.07	131 0.15 8.88 2.64	819 0.94 55.53 1.43	1475 1.69
AZ	18 0.02 1.40 1.10	197 0.23 15.37 0.83	100 0.11 7.80 2.02	967 1.11 75.43 1.69	1282 1.47
CA	180 0.21 2.83 10.98	643 0.74 10.13 2.72	359 0.41 5.65 7.24	5168 5.92 81.39 9.04	6350 7.27
CO	36 0.04 2.33 2.20	570 0.65 36.96 2.41	36 0.04 2.33 0.73	900 1.03 58.37 1.57	1542 1.76
CT	18 0.02 3.33 1.10	0 0.00 0.00 0.00	36 0.04 6.67 0.73	486 0.56 90.00 0.85	540 0.62

Data Exploration

Best Practices

BEST PRACTICE

- Avoid Categorical Data as Numbers – If You Can
- Address Missing Data Appropriately
- Thoroughly Clean the Data!
- Data is More Than Just Numbers and Text
  - Data is: People, Animals, Plants, Environment, AI
  - Data is a snapshot of information, not the whole picture
- Treat Data Like a 3-D Living Thing
  - Use different perspectives (wise men and the elephant)
  - Consider what the data is not telling you
  - Consider the age of the data
- S.W.O.T. the Data
  - Strengths
  - Limitations/Weaknesses
  - Opportunities
  - Threats
- Always review your log

Data Driven Modeling

WHO Global Table Data

Data Driven Modeling

Key Considerations

- 1) Identify the roles of your variables – what are the variable relationships
  - Predictors (IV)
  - Outcomes (DV)
  - Confounders
  - Covariates
- 2) Variable roles determine their location in the research question
- 3) Research question structure informs the analysis type – Next Section

Outcome = Predictor1 + Predictor2 + Confounding + Covariate

DV = IV1 + IV2 + Confounding + Covariate

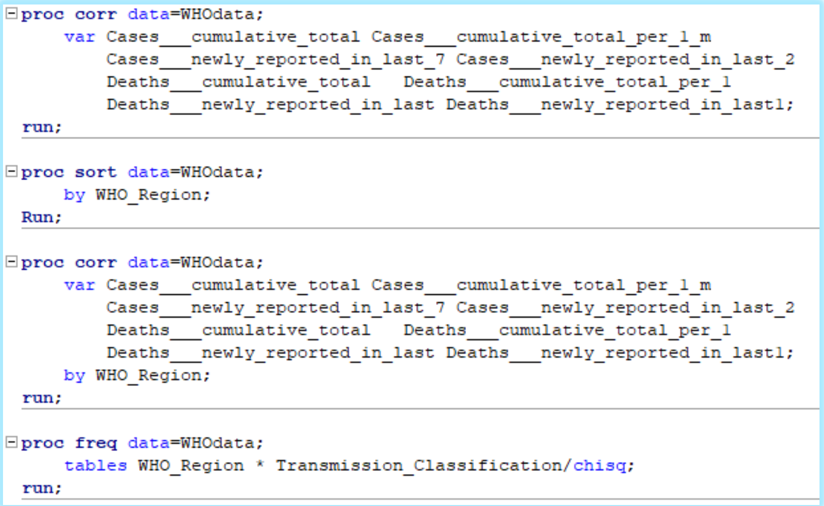
Data Driven Modeling

SAS Procedures

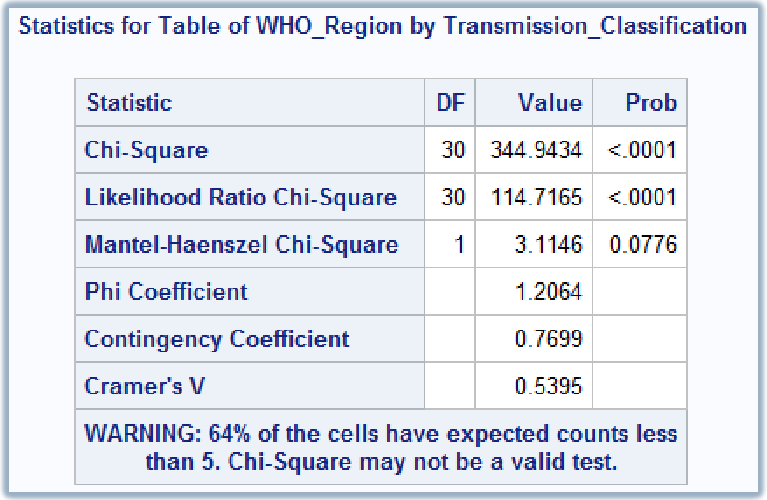
- SAS Procedures
  - Proc Freq
  - Proc Means
  - Proc Univariate
  - Proc Corr
- The Data Step
  - Implement adjustments to the data
- Additional Helpful Procedures
  - Proc Contents
  - Proc Sort
  - Proc SQL
  - Proc Print
  - SAS Macros

Data Driven Modeling

COVID Example



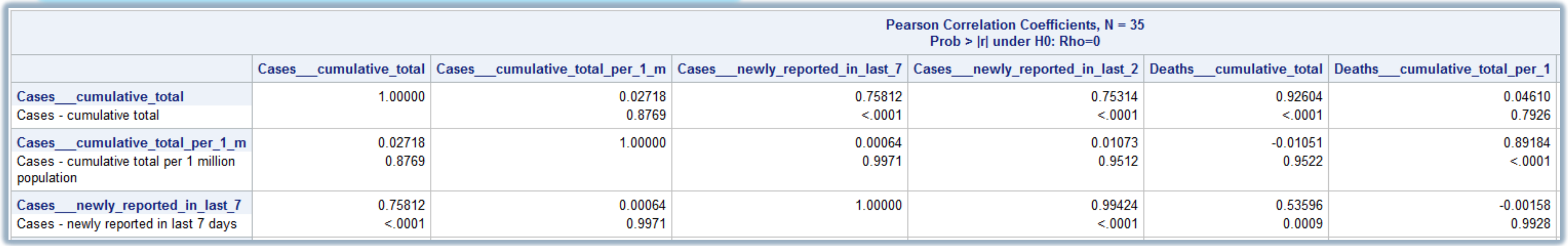
```
proc corr data=WHOdata;
    var Cases___cumulative_total Cases___cumulative_total_per_1_m
        Cases___newly_reported_in_last_7 Cases___newly_reported_in_last_2
        Deaths___cumulative_total   Deaths___cumulative_total_per_1
        Deaths___newly_reported_in_last Deaths___newly_reported_in_last1;
run;

proc sort data=WHOdata;
    by WHO_Region;
Run;

proc corr data=WHOdata;
    var Cases___cumulative_total Cases___cumulative_total_per_1_m
        Cases___newly_reported_in_last_7 Cases___newly_reported_in_last_2
        Deaths___cumulative_total   Deaths___cumulative_total_per_1
        Deaths___newly_reported_in_last Deaths___newly_reported_in_last1;
    by WHO_Region;
run;

proc freq data=WHOdata;
    tables WHO_Region * Transmission_Classification/chisq;
run;
```



Statistics for Table of WHO_Region by Transmission_Classification

Statistic	DF	Value	Prob
Chi-Square	30	344.9434	<.0001
Likelihood Ratio Chi-Square	30	114.7165	<.0001
Mantel-Haenszel Chi-Square	1	3.1146	0.0776
Phi Coefficient		1.2064	
Contingency Coefficient		0.7699	
Cramer's V		0.5395	

WARNING: 64% of the cells have expected counts less than 5. Chi-Square may not be a valid test.



Pearson Correlation Coefficients, N = 35
Prob > |r| under H0: Rho=0

	Cases___cumulative_total	Cases___cumulative_total_per_1_m	Cases___newly_reported_in_last_7	Cases___newly_reported_in_last_2	Deaths___cumulative_total	Deaths___cumulative_total_per_1
Cases___cumulative_total Cases - cumulative total	1.00000	0.02718 0.8769	0.75812 <.0001	0.75314 <.0001	0.92604 <.0001	0.04610 0.7926
Cases___cumulative_total_per_1_m Cases - cumulative total per 1 million population	0.02718 0.8769	1.00000	0.00064 0.9971	0.01073 0.9512	-0.01051 0.9522	0.89184 <.0001
Cases___newly_reported_in_last_7 Cases - newly reported in last 7 days	0.75812 <.0001	0.00064 0.9971	1.00000	0.99424 <.0001	0.53596 0.0009	-0.00158 0.9928

#SASGF

SAS® GLOBAL FORUM 2021

Data Driven Modeling

Best Practices

BEST PRACTICE

- What Has Been Done?
  - Check the work of others for guidance on variable relationships
  - Address the findings of past research
- Variable Couples Counseling
  - Even if you think a variable is not related to another, check anyways
  - Pay attention to the impact one variable may have on the relationships of others
- Data Structure Incompatibility – Mathematical Theory
  - Consider differences between numeric/categorical data
    - Potential use of binning
  - Consider limitations of mixing within-group data structures
    - Nominal & Ordinal, Multi-Level & Binary/Dichotomous
    - Interval & Ratio, Discrete vs Continuous
- Always review your log

Matching Your Question to a Model

OpenICPSR

COVID Isolation on Sleep and Health in Healthcare Workers

Matching Your Question to a Model

Key Considerations

- Checking Model Assumptions – Common Assumptions
  - Normality
  - Homogeneity of Variance
  - Homogeneity of Variance-Covariance Matrices
  - Linear Relationships
  - Absence of Multicollinearity
  - Absence of Auto-Correlation
  - Randomization
  - Large Sample Size
- Model Assumption Violations
  - Can sometimes be mitigated through variable adjustment
  - Fatal violations require a change in model choice

Matching Your Question to a Model

SAS Procedures

- Normality
  - Proc Univariate
  - Proc Capability
- Homoscedasticity
  - Proc GLM
  - Proc Reg
  - Proc Model
  - Proc Transreg

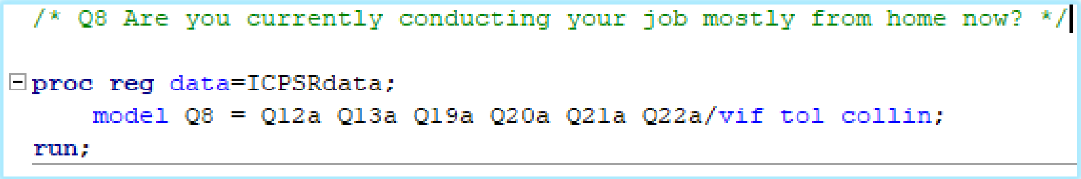
- Homogeneity of V-C Matrices
  - Proc Discrim
  - Proc GLM
  - Proc Standard
- Multicollinearity
  - Proc Corr
  - Proc Reg

- Autocorrelation
  - Proc Reg
  - Proc Autoreg
- Linear Relationship
  - Proc Reg
  - Proc Corr
  - Proc Logistic

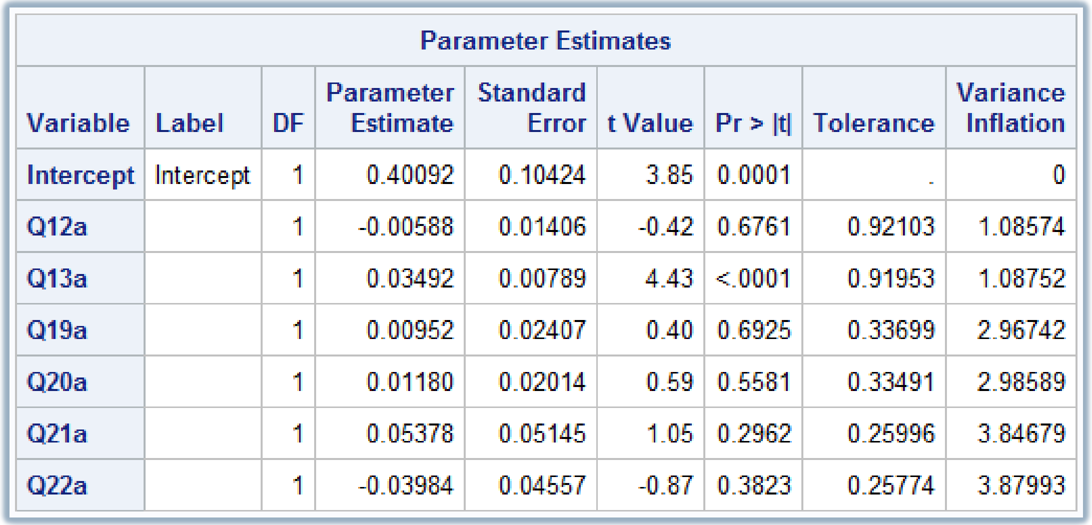
Note: Some tests require multiple steps across different SAS procedures

Matching Your Question to a Model

COVID Example

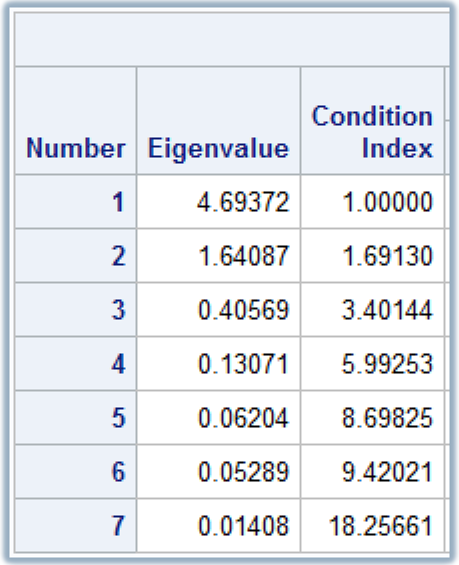


```
 /* Q8 Are you currently conducting your job mostly from home now? */

proc reg data=ICPSRdata;
     model Q8 = Q12a Q13a Q19a Q20a Q21a Q22a/vif tol collin;
 run;
```



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.40092	0.10424	3.85	0.0001	.	0
Q12a		1	-0.00588	0.01406	-0.42	0.6761	0.92103	1.08574
Q13a		1	0.03492	0.00789	4.43	<.0001	0.91953	1.08752
Q19a		1	0.00952	0.02407	0.40	0.6925	0.33699	2.96742
Q20a		1	0.01180	0.02014	0.59	0.5581	0.33491	2.98589
Q21a		1	0.05378	0.05145	1.05	0.2962	0.25996	3.84679
Q22a		1	-0.03984	0.04557	-0.87	0.3823	0.25774	3.87993



Number	Eigenvalue	Condition Index
1	4.69372	1.00000
2	1.64087	1.69130
3	0.40569	3.40144
4	0.13071	5.99253
5	0.06204	8.69825
6	0.05289	9.42021
7	0.01408	18.25661

Matching Your Question to a Model

Best Practices

BEST PRACTICE

- More Than One Way…
  - There are numerous routes to test an assumption
    - Use multiple or narrow in on the most appropriate
- Do Not Hesitate to Switch Models If Needed
- Do Not Force a Model
- The High-Rollers Club
  - The more complex the analysis, generally the more numerous and complex the assumptions
  - Violations/naïve analyses can be very harmful
  - Enlist help
- Null Results Are NOT Necessarily a Model Failure
- Always review your log

Evaluate Your Model

Illinois COVID Vaccination Distribution

Evaluate Your Model

Key Considerations

- After The Model is Run – You Are Not Done!
- Check for Key Model Health Indicators
  - Predictive Power
  - Model Fit
- Consider/Check Data Health Indicators
  - Validity
  - Reliability
  - Generalizability

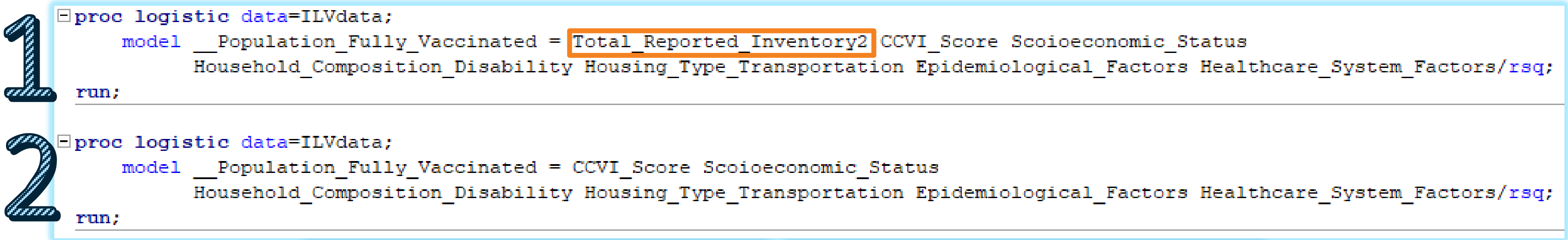
Evaluate Your Model

SAS Procedures

- Power
  - Cox-Snell: Proc Logistic & Proc Reg
  - Tjur: Proc Logistic & Proc Ttest
- Model Fit
  - Pearson: Proc Logistic
  - Hosmer-Lemeshow: Proc Logistic
  - Stukel: Proc Logistic
  - %goflogit macro
  - AIC, etc: Proc Phreg, Proc Reg, & Proc Logistic

Evaluate Your Model

COVID Example



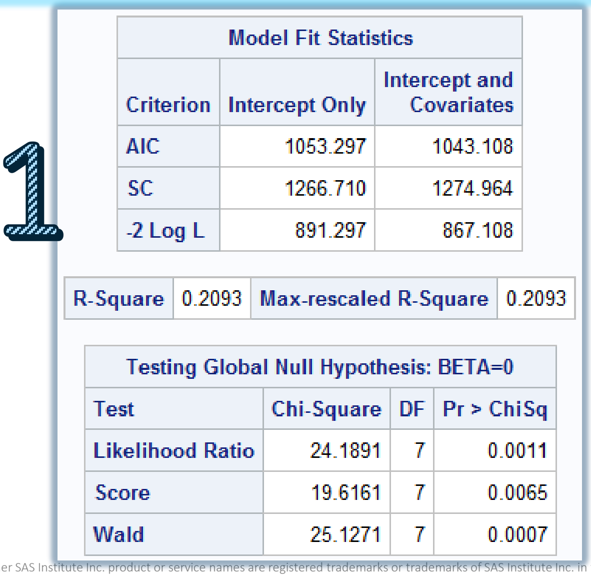
1

```
proc logistic data=ILVdata;
    model __Population_Fully_Vaccinated = Total_Reported_Inventory2 CCVI_Score Scoioeconomic_Status
          Household_Composition_Disability Housing_Type_Transportation Epidemiological_Factors Healthcare_System_Factors/rsq;
run;
```

2

```
proc logistic data=ILVdata;
    model __Population_Fully_Vaccinated = CCVI_Score Scoioeconomic_Status
          Household_Composition_Disability Housing_Type_Transportation Epidemiological_Factors Healthcare_System_Factors/rsq;
run;
```

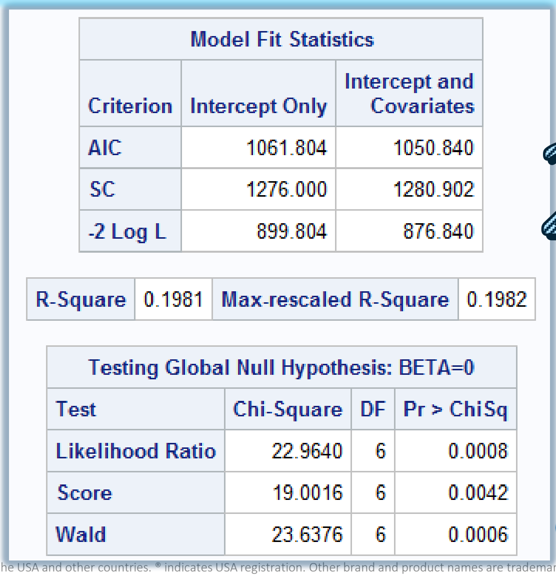
1



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1053.297	1043.108
SC	1266.710	1274.964
-2 Log L	891.297	867.108

R-Square	0.2093	Max-rescaled R-Square	0.2093

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.1891	7	0.0011
Score	19.6161	7	0.0065
Wald	25.1271	7	0.0007



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1061.804	1050.840
SC	1276.000	1280.902
-2 Log L	899.804	876.840

R-Square	0.1981	Max-rescaled R-Square	0.1982

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22.9640	6	0.0008
Score	19.0016	6	0.0042
Wald	23.6376	6	0.0006

2

Evaluate Your Model

Best Practices

BEST PRACTICE

- More Than One Way…
  - There are numerous routes to test power and model fit
    - Use multiple or narrow in on the most appropriate
- Do Not Hesitate to Switch Models If Needed
- Do Not Hesitate to Restructure a Model in Poor Health
- Do Not Force a Model
- Null Results Are NOT Necessarily a Model Failure
- Always review your log

Conclusion

Conclusion

Key Takeaways

- Choosing and Importing Data
  - Choose/collect data that matches your question
  - Consider research method basics
  - Pay attention to data structure, size, and generalizability.
- Data Exploration
  - Get to know your data! How to run basic descriptive statistics with consideration to data type.
- Data Driven Modeling
  - Identify predictors (IV), outcomes (DV), confounders, and covariates
- Matching Your Question to a Model
  - Make sure your model assumptions fit your question and data. Every model has its own set of assumptions! Violation of these assumptions lead to incorrect conclusions
- Evaluate Your Model
  - Check and refine your model performance through exploration of power and model fit
  - If necessary, evaluate validity, reliability, and generalizability of data

Conclusion

Best Practices

BEST PRACTICE

- Choosing and Importing Data
  - Pay attention to where your data is coming from & know that data has limitations
  - Practice good data storage basics, maintain an untouched original dataset, & document adjustments.
- Data Exploration
  - Address missing data appropriately & avoid categorical data as numbers
  - See the face of data & know that data is a living 3-dimensional entity
- Data Driven Modeling
  - Consider data structure incompatibility & Test variable relationships
  - Document and implement findings from past research
- Matching Your Question to a Model
  - There are more than one way to test assumptions – use them
  - Do not hesitate to switch models, do not force a model
  - Consider model complexity
- Evaluate Your Model
  - There is more than one way to test power and model fit – use them
  - Do not hesitate to appropriately restructure a model in poor health
  - Null results do not mean model failure/incompatibility.
- Always review your log

Conclusion

Remember

- Data is everywhere and understanding data science is a growing necessity for navigating today’s world.
- This journey should not be done solo. Interdisciplinary teams of scientists/researchers, statisticians, programmers, and advocates/specialists are needed to make the most of the information available to us.
- Having an understanding of the analytic process will help create the bridge of communication needed to answer the complex questions of today.

Resources

Further Reading

- A Gentle Introduction to Statistics Using SAS Studio – book
- Introduction to Biostatistics with JMP – book
- Fundamentals of Programming in SAS – book
- Practical Data Analysis with JMP – book
- Real World Health Care Data Analysis Causal Methods and Implementation Using SAS – book
- Lexjansen.com – SAS Papers

Thank you!

Contact Information
d.n.schreibergregory@gmail.com

VIRTUAL

SAS® GLOBAL FORUM 2021

#SASGF