SESUG Paper 028-2021

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

Deanna Schreiber-Gregory, Juxdapoze, LLC

ABSTRACT

Data Science is the new Space Race, launching us into a world of immeasurable possibility, but with only a few people to help us navigate it. As we dig deeper, discover more, and risk more, we can be simultaneously led to both great insight and loss. If we do not know what we are doing, Data Science can be a very dangerous thing. It is important for us all to learn at least a little bit about the possibilities and risks of this field of study, so we can navigate it together.

This paper was written to give individuals new to SAS® and/or Analytics a gentle nudge in the direction of the possibilities available through Data Science and SAS. It is designed to help you navigate through the process of data exploration by using publicly available COVID 19 data. We have all seen how fragile this data reporting can be, and this paper uses this fragility to help explain the dangers of an inappropriately implemented analytic process. Together, we will briefly touch on current best practices and common errors that occur at the different steps of the analytic process (choosing data, exploring data, building and running a model, checking and refining model performance) while simultaneously reviewing common SAS procedures used in each of these steps (Data Step, Univariate Procedures, Multivariate Procedures, Power & Model Fit Procedures). At the end of this paper, the author provides several citations and recommended readings to help interested analysts further their education in Data Science implementation.

Data is everywhere and understanding data science is a growing necessity for navigating today's world. This paper is meant to help give individuals a snapshot of insight into the vastness of possibility that is Data Science.

INTRODUCTION

Data Science is the new Space Race, launching us into a world of immeasurable possibility, but with only a few people to help us navigate it. As we dig deeper, discover more, and risk more, we can be simultaneously led to both great insight and loss. If we do not know what we are doing, Data Science can be a very dangerous thing. It is important for us all to learn at least a little bit about the possibilities and risks of this field of study, so we can navigate it together.

WHAT IS DATA SCIENCE?

According to the Merriam-Webster dictionary, data science is….not in it. As of March, 2021, the term "data science" does not have a definition recorded in one of the most famous and well-referenced dictionaries of the English language. Given the term's popularity and the fact that its origins can be traced back to the year 2001, it is a wonder why it doesn't have a formal definition recorded. This is probably because the term represents an idea that is still evolving. In all honesty, the domain of data science is barely through childhood. How could we assign a career definition to such a young field of study?

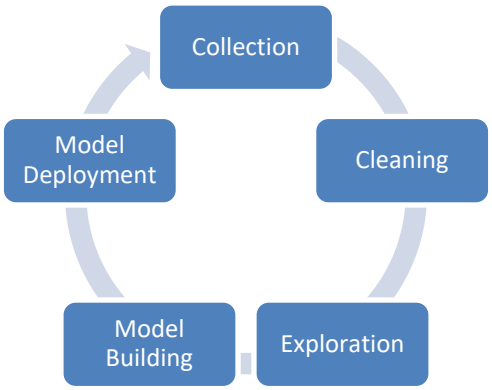
Even though we may not be able to confidently point to the definition of data science right now, there is one thing we do know: we NEED it. How do we know we need it? Because everyone tells us we do. How do they know? Because that is what the analysts say. How do the analysts know?...I'm sure you can see where this is going.

The thing is, even though data science is a young study, just coming into its own personality and identity, it is a real study that encompasses knowledge, experience, and facts as old as life itself. Data science is founded in mathematics and observation and communicated through the interdisciplinary discussion of research and advancement. Data science is something we have been doing for centuries. We collect data, we figure out what the data means, and we act on it. This process is usually done by teams instead of a single individual. Before data science was coined and started gaining ground, individuals known as

"unicorns" in the science world started to appear. These individuals could not only collect, program, and analyze the data using advanced mathematics, but they could also confidently interpret the meaning of the results given their own expertise in the focus field. Later, these "unicorns" took on a different name: data scientist.

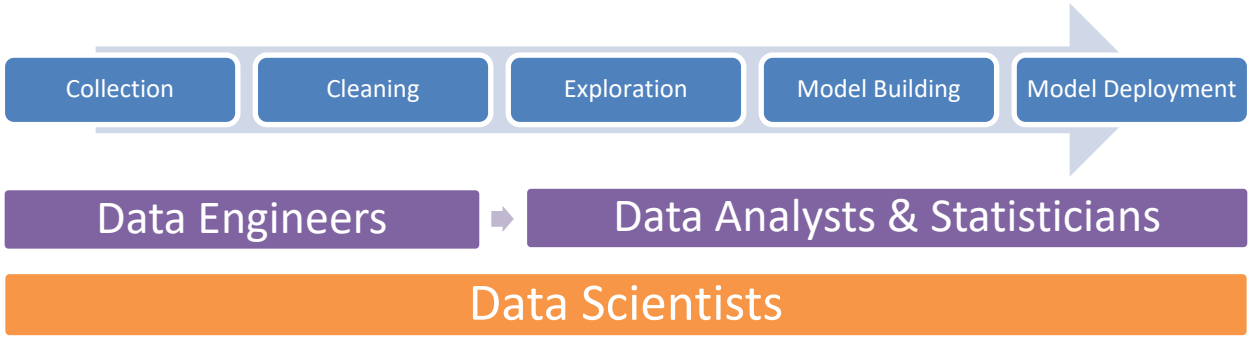
THE ANALYTIC PROCESS AND DATA SCIENCE

At the core of data science is the analytic process. Oftentimes thought of as linear, the analytic process can be described in a variety of different ways using any number of steps, but is most often displayed in a cyclical manner.



The reason for this cyclical representation of a process otherwise thought of as linear, is that a true analysis is designed to meaningfully feed back into itself. Analyses are designed with the intent to inform and refine their own methodology. This is what makes the common (linear) understanding of the analytic process so error-prone and messy, and the data science (cyclical) approach even more appealing.

When multi-disciplinary research teams were first being formed, the analytic process was split between two main groups of individuals (depending on your field, this number could be higher, but we are going to approach this from a basic standpoint for now). These groups of individuals are commonly referred to as data engineers and data analysts/statisticians. Each group has their separate area of expertise. Data engineers are versed in data theory, computer science, missing data methodology, and data structure. Data analysts and statisticians are versed in statistical theory and mathematics, have a familiarity with data theory and missing data methodology, and may have some expertise in the topic area being explored. The communication between these two groups of individuals is key, as the work of either group must be incorporated into the understanding and work of the other. Thought of as ideal for a linear analytic process, the separation of expertise and knowledge between these two groups can be problematic when considering the natural cyclical nature of research and analysis. In response to this, data scientists, as discussed earlier, must be versed in all of the above, and therefore, can seamlessly incorporate the knowledge, experience, and best practices of one into the other.



This cyclical structure, and data science's role in it, is further supported when you consider the scientific method/process on the whole. As any student researcher can explain, the scientific process is composed of these steps:

Do we stop at report? No, we do not. Just because a study was completed, we never take the results as indisputable fact. After all, when a statistical test reveals insignificant results, we say "the test failed to reject the null hypothesis", and if the test was significant, we say "the test rejected the null hypothesis". Notice the distinct and glaring absence of "prove". We have proven nothing. We have only furthered our knowledge with the intent that it will be fed back into the loop to inform future analyses. Proof is for mathematics, progress is for analytics.

THE COVID PUZZLE

As the year 2020 unfolded, we were quickly met with the urgency of a rising pandemic. Once the emergency was declared, politicians, health professionals, financial institutions, education administrators, and everyone else turned to data and analytics to make sense of what was happening. We expected to find cold hard facts to guide and calm us. Unfortunately, what we found was a fragile, inconsistent, and error-prone process that consistently left us both confused and unsure of what to do next. Where mathematics had historically given us direction, we felt abandoned in an unknown landscape.

So, what happened?

As discussed numerous times by a variety of professionals, several assumptions had been made that did not fit the situation we had found ourselves in. Data was provided in real-time and analyses were expected within very tight, and virtually impossible, windows of time. Data, itself, was inconsistent in its reliability, form, and source. Reporting expectations were also inconsistent, as there had been no time to derive standards that the analytic field could follow. The analyses and reporting had to basically be done blind, as researchers and field professionals had little to no prior knowledge to base their work off of. Given all of this, one can see how the natural feedback loop became distorted as the ethics and theoretical practice of data science became more stressed.

This brings us to where we are now, exposed to the dangers of naïve and fragile analytics. Through COVID, we have all seen how fragile data analysis and reporting can be, and we must now use this newfound

knowledge to avoid the dangers of an inappropriately implemented analytic process in the future. The data provided within this paper has been gathered from public sources and readers are encouraged to apply what they learn to these datasets to further their understanding of analytic application.

In the following sections we will briefly touch on current best practices and common errors that occur at the different steps of the analytic process (choosing data, exploring data, building and running a model, checking and refining model performance) while simultaneously reviewing common SAS® procedures used in each of these steps (Data Step, Univariate Procedures, Multivariate Procedures, Power & Model Fit Procedures). This paper is by no means a comprehensive introduction to the analytic process. Instead, it is a basic topographical map to help give the reader a "satellite image" of what to expect. Resources for further exploration are provided and encouraged to help equip the analytic adventurer with the tools necessary to navigate the world of data within the depths of cyberspace.

STEP 1: COLLECTION - CHOOSING & IMPORTING DATA

Let us again consider the Scientific Method. In review of the steps, you may notice that the necessity and appearance of data does not occur until the fifth step…of eight! Throughout the steps of Observe, Question, Research, and Hypothesize, the topic of gathering actual data for research is not touched. Not until Experiment would data used in the study finally appear. Considering its late arrival, why do we put so much emphasis on it?

No project has ever found results without data. Whether the data is qualitative, quantitative, theoretical, or observational, the best researchers use gathered evidence to support their case. Therefore, it is only natural that the start of any project lies in the research and identification of *appropriate* data. The data does not appear until later, because the first four steps are dedicated solely to identifying *what* the data actually is and what it should look like.

KEY CONSIDERATIONS

Key 1: Choose/collect data that matches your question.

It is generally understood that there are two main routes through which data is gathered and analyzed: Primary and Secondary. Through Primary Analyses, the data is gathered with a specific research question or set of questions in mind. This data is hand-picked, structured, gathered, and refined according to these questions. Assumptions and bias from the population are considered immediately and addressed before the first data point is recorded. The resulting dataset is therefore bespoke to the parent study. Secondary Analyses, on the other hand, use previously gathered data to answer a question. Data for these types of analyses are usually recorded before a particular project has even entered the "Observe" step of the Scientific Method. Though this type of analyses may seem inferior to Primary Analyses on paper, it can be just as powerful, as long as the data fits the question.

Another way to divide data types is by the research question, itself. In Experimental Research, researchers introduce a planned intervention and study its affects. Experimental studies are almost always Primary Analyses by nature (at least for the main groups), and usually incorporate assumptions of randomization and generalizability from the start. Randomized Controlled Trials (RCTs) are a very common type of Experimental Research. On the other hand, in Observational Research, researchers observe the effect of a risk factor, test, treatment, or other intervention without controlling who in their participant pool is affected. Common types of Observational Research include Cohort and Case Control studies. Like Primary and Secondary analyses, both types of research have strengths and weaknesses.

Data can also be gathered in a wide variety of different formats. Through surveys, clinical records, neutral or biased observation, or even mined from extremely large data sources (ex. Social media), the source and format through which data was gathered has a direct impact on the quality, relationship, and performance of that data within a model.

No one method holds absolute authority over the other. It all depends on the question you are trying to answer. It is important to very seriously consider the question you are trying to answer and identify to which analytic type your question falls. If you choose the wrong environment to gather your data, your results will be meaningless and difficult to defend.

Key 2: Consider research method basics.

Do not disregard the Scientific Method. Take each step seriously, as there is a reason it was identified and included in the widely accepted order of conducting research. By approaching each step with care and diligence, your identification of the appropriate dataset will become all the clearer.

Key 3: Pay attention to data structure, size, and generalizability.

If you want to maximize the impact of your research, you need to maximize the representation of your data. Many research models have restrictions on sample size, data structure, and composition. In order to use the most appropriate model, your data pool must be adequately sized with appropriate representation from your target groups. Too much missing data or lopsided groupings can inappropriately weight an analysis, leading to skewed results.

Key 4: Make note of data that needs to be merged and in what format the data is stored.

Data comes in a variety of formats. Excel, SPSS, R, SAS, and text, all store data in different ways. It is important to identify how that data is being stored and to perform any necessary transformations on the data to make sure it is read appropriately into the analytic software of your choice (ex. SAS). Just because it looks right in the original dataset, doesn't mean it was read right by the software. If the data needs to be merged, this step becomes even more crucial. The same data source could store an iteration of the data in one format, but completely change the nature, order, and formatting of the data in another. It is important to identify as many possible pitfalls that could affect data import up front, so that they can be mitigated as soon as possible. It is also important that quality checks are performed to help identify any mistakes that you may have missed.

SAS PROCEDURES

As far as getting data into SAS, there are numerous ways you can do this: 1) enter it directly using data step, 2) use the IMPORT procedure, 3) point-and-click File (or New) -> Import data and follow the prompts. For this paper, the IMPORT procedure was used:

```
/* CDC COVID Tracker Data */


PROC IMPORT OUT= cdcdata_raw

            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=cdcdata_raw (obs=200);

run;


data cdcdata;
```

```
        set cdcdata_raw;

        if hosp_yn = 'Missing' then hosp_yn = 'Unknown';
        if icu_yn = 'Missing' then icu_yn = 'Unknown';
        if death_yn = 'Missing' then death_yn = 'Unknown';
        if medcond_yn = 'Missing' then medcond_yn = 'Unknown';

        length days 8.;
               days = cdc_report_dt - cdc_case_earliest_dt;
run;
```

Once you have decided how to import your data, you must also decide how the data will be stored while you work on it. You can store the imported data file as a permanent or temporary file. I usually store mine as temporary, as this allows me to reuse my code for future projects without too much adjustment, and allows me to maintain a documentation stream that does not affect the original dataset, but shows the proposed adjustments.

Once you have completed the necessary adjustments to the data, you can export it into a different format to share with colleagues if necessary. I use the EXPORT procedure to do this.

```
PROC EXPORT
  DATA=cdcdata
  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;
```

This allows me to collaborate with colleagues who are not otherwise familiar with SAS, or who do not have SAS on their machines, but need to look at the data.

BEST PRACTICES

There are several best practices that can be employed, even in the choosing and importing data phase.

Practice 1: Pay attention to sources.

Pay attention to where your data is coming from. Is the data source a reliable one? How did they collect it? What format does it come in? How old is it? Who owns the data?

Practice 2: Understand that data has limitations.

There is only so much information stored in one data source. Acknowledge that any one data set may not have all of the information you need. You may have to merge datasets, or consider the limitations of a set's information in your analysis. Datasets are finite, even the very big ones. Just because the title and general layout of the data matches your question, doesn't mean that the actual observations will. You still need to go through every step of the analytic process to make sure that your data is appropriate to answer your question.

Practice 3: Practice good data import and processing basics.

Make sure there is no excess information at the beginning or end of your dataset (you can do this by specifying where the data begins or ends). Make sure that your variable names make sense. Make use of labels and formats, they are your friends and will make interpreting tables so much easier. Use labels instead of long names as excessively long names before import will get shortened by SAS and could quickly become difficult to interpret.

Practice 4: Maintain an untouched original dataset without adjustments.

There are two big benefits to code: 1) you have written documentation of your work which can be easily used to update analyses or replicate a study, and 2) you end up with detailed documentation as to everything that has been adjusted to your dataset. By maintaining an untouched original dataset, you can back-trace everything and correct mistakes quickly and easily. This documentation is also necessary for any auditing of the analysis and adjustments that may come up later on in a study.

Practice 5: Comment your code.

Make liberal use of the ability to comment in your code. Why did you choose these options when importing your data? Where did this adjustment come from? When did you do it (use dates)? If more than one person is working on a code, make sure to initial your comments so that your colleagues can understand what you are doing.

Practice 6: Always review your log.

SAS's log does a fantastic job of notifying you when it runs into an error or if the data is toeing the line of a procedure's underlying methodology. The log is not exhaustive and does not catch everything, but is a great first step in making sure everything ran the way it should.

STEP 2: CLEANING & DESCRIBING - DATA EXPLORATION

Now that you have your data into SAS, get to know it! Before we can start putting our data into a model, we must become deeply familiar with its structure. It is important to lay out the various types of data that we have and acknowledge/account for how these different types will interact with each other.

KEY CONSIDERATIONS

Data falls into two main mutually exclusive classification types, numeric and categorical. Within these classification types are sub classifications that further break down how a particular variable is described and handled.

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

Key 1: This is one of the MOST important steps.

This is a natural part of data exploration. It can be arduous and a bit mind-numbing at times, but skipping this step WILL very negatively impact your final model. It is never recommended to approach this step half-heartedly. Know that it will take time, but also know that the time is well spent, as you will be more confident in your model building and assessment later on.

Key 2: Consider the types of numeric variables.

A key aspect about numeric variables is that you can confidently and objectively measure the distance from one value to the next. Numeric variables can be described as interval, ratio, discrete, or continuous.

Interval variables are those consisting of interval data. Interval data is measured on a scale along the whole of which intervals are equal. This data type has no true zero.

Ratio variables are a type of interval variable with the additional property that ratios are meaningful. This data type has a true 0, and doubling a particular value results in a true double.

Discrete variables can only take on certain values (usually whole numbers) within a specified scale. There is no true decimal (ex: you can not have 1.2 children).

Continuous variables can be measured to any level of precision.

Key 3: Consider the types of categorical variables.

A categorical variable is any variable made up of categories of objects or ideas. Categorical variables can be described as binary/dichotomous, multi-level, nominal, ordinal, or dummy.

A binary or dichotomous variable is a categorical variable that has two mutually exclusive categories. Though generally thought of as the same, a binary variable represents a presence/absence scenario (Yes or No), while a dichotomous variable represents a two group category (Male or Female).

A multi-level variable is one that has more than two levels. The commonly used “Likert Scale” structure falls into this category.

A nominal variable is one in which the categories do not adhere to any organization. There is no level or order in these variables, they are simply categories (or names, identifications, etc).

An ordinal variable is one in which the categories adhere to some order.

A dummy variable is created as a way to recode categorical variables that have more than two categories into a series of variables, all of which follow a binary structure (absence/presence) and usually take on values of 0 or 1. There are seven recorded steps to the creation of dummy variables which is beyond the scope of this paper. However, their use is beneficial in some models, and it is important to understand that they are, indeed, categorical variables.

Key 4: Build tables that describe your data and population.

In this step, it is important to build a fair number of tables that can be used to describe your population. If you are writing a research paper, these tables will be very useful when writing your methods section and the beginning of your results section. For categorical variables, what is the frequency of value occurrence? Is there a significant difference in value distribution across your target variable(s)? For numeric variables, what is your min, max, mean, median, standard deviation, and histrogram distribution? Do you have any outliers that need to be looked into further?

Key 5: Explore missing values.

Do not overlook the incidence of missing values. If you have missing values, there are a variety of ways to handle their existence (or literal lack thereof). Before considering any type of imputation or acceptance of data loss, it is important to explore the reasons as to why this data is missing. The existence of missing data is, in and of itself, data, as there is always a reason to the missingness. Finding this reason will help guide how you handle it.

SAS PROCEDURES

There are numerous SAS procedures that can be employed for data exploration. No matter your data type or distribution, there is a exists a procedure to handle it.

Commonly used procedures include FREQ, MEANS, UNIVARIATE, and CORR. It is important to familiarize yourself with these procedures and the different output and options they offer.

```
/* Health.gov Data */


PROC IMPORT OUT= healthgov_raw

            DATAFILE= "D:\[Conference] Current Papers\Back to Basics\Resources - 
Data\Health     Data.gov     Data\reported_hospital_capacity_admissions_facility-
level_weekly_average_timeseries_20201207.xlsx"

            DBMS=XLSX      REPLACE;      SHEET='reported_hospital_capacity_admi'; 
GETNAMES=YES;

RUN;


data healthgov;

       set healthgov_raw;

run;


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;


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

During exploration, if you find necessity to adjust a variable to better fit your model, or clean up a bit of messiness, the DATA step can be employed. You can easily adjust your data, rename variables, or create new ones. These adjustments are not only carried through in an easy-to-understand manner by SAS, but the existence of these adjustments within your code serve as archived documentation to how your data is being adjusted. This documentation will be of use when writing about your results or responding to an audit on your research.

There are plenty of other helpful procedures in SAS that are equipped to assist you during this step. Some popular ones are: CONTENTS, SORT, SQL, PRINT, and SAS Macro processing.

BEST PRACTICES

Best practices in data cleaning and exploration is a popular topic that has been written about for years. Here, I will cover some practices that I have found particularly useful to reduce headache in this stage.

Practice 1: Address missing data appropriately.

As stated in the Key Considerations subsection, it is important to appropriately approach missing data. This aspect of data modeling can be easy to overlook. It can be frustrating and difficult to understand. However, its existence and handling is an important one. Do not disregard the impact and meaning of missing data.

Practice 2: Avoid categorical data as numbers.

This one could be debated, but I have found that if you avoid storing categorical data as numbers, you are less likely to give in to the temptation of treating them as numbers. Multi-level ordinal variables are especially susceptible to this, as they are often thought to be more interval in nature. They are absolutely not numeric variables, so numeric procedures would apply assumptions upon them that would be inappropriate. For simple data exploration, the employment of a numeric procedure on a variable such as this could be insightful (ex: average rating), but to place a categorical variable into a model that will see it as numeric, can invite error into your results. I recommend placing this limitation on your data so that these variables are treated with the appropriate assumptions when building and employing your model.

Practice 3: See the face of data.

Data is more than just numbers and text. Data is: people, animals, plants, environment, artificial intelligence, or ideas. Data is a snapshot of information, not the whole picture. When working with data, you are working with a living, breathing entity that ages, grows, and can be easily misunderstood. You are not just working with numbers, you are working with entities that exist within our own environments. While considering this, it is worthwhile to look at data

from different angles and perspectives (consider the parable about the blind/wise men and an elephant). Think about what the data is NOT telling you.

Practice 4: Apply S.W.O.T. methods to data exploration.

A SWOT analysis is used to assess four key aspects of an entity. Normally used in business, it still has a place in data exploration. SWOT stands for strengths, weaknesses, opportunities, and threats. Identifying the strengths and weaknesses of data seems pretty commonplace, but opportunities and threats? Think of it this way, strengths and weaknesses are current, while opportunities and threats are potential. So if you can identify the current strengths and weaknesses of a dataset in relation to your project, what about the things the data is not telling you? What else could this data address? What is missing? If an event occurred, how would it impact the data structure? Answering questions like these will open up another door to identifying the strengths and weaknesses of a dataset, as it will give you insight into the potential fragility or stability of what you have found, which will aid in the discussion of the results.

Practice 5: Always review your log.

SAS's log does a fantastic job of notifying you when it runs into an error or if the data is toeing the line of a procedure's underlying methodology. The log is not exhaustive and does not catch everything, but is a great first step in making sure everything ran the way it should. If you think you read this before, you did, this is how important it is.

STEP 3: MODEL BUILDING - DATA DRIVEN MODELING

So you have a pretty firm handle on your data, let's start putting it into a model!

KEY CONSIDERATIONS

Key 1: Identify predictors (IV), outcomes (DV), confounders, and covariates.

At this point, you should already have at least a vaguely structured research question in your mind, now is the time to formally map out that question and how your data addresses it.

One of the most talked about variables in your project will be your **dependent variable (DV)** or **outcome variable**. In experimental research, an independent variable is the one that is not manipulated by the experimenter. Its value depends on the variables that are being manipulated. In observational research, an outcome variable is one whose value is being predicted by one or more predictor variables.

The other high-profile variable in your study will be your **independent variable (IV)** or **predictor variable**. In experimental research, the independent variable is one that is manipulated by the experimenter. Its value does not depend on any other variables. In observational research, the predictor variable is one the target variable used to try to predict values of the outcome variable.

The IV/predictor and DV/outcome variables represent the most significant relationship being explored in your model. Though the above two variable types will probably be the most referenced in your study, there are a few other types that need to be considered in order to build the best fit model.

A **covariate** is a variable that has at least the potential to have a relationships with the outcome variable. Covariates are included in the model because of their relationship to the outcome. Their relationship with other predictor variables needs to be tested, as an argument or significant relationship between a covariate and a predictor can be very bad (ie. collinearity/codependence). Given that covariates still exist on the "independent" side of the equal sign (right side, needs to be independent from the other predictors), any relationship with a variable other than the outcome, will need to be handled appropriately. Oftentimes, covariates have the potential to drop out of the model, due to non-significance.

A **confounding** variable is a variable other than the predictor variable that could potentially affect the outcome variable. This variable may or may not have been measured in your study. It could be derived

from a relationship between two variables, or should at least be talked about as a confounding factor that could account for some loss in predictive power. These variable types are allowed inclusion in a model, because their impact focuses on how the predictor relates to the outcome. It does not need to have a direct relationship with the outcome, though it sometimes does. In essence, I like to think of confounding variables as language translators. They help a specific predictor variable communicate appropriately with the model. However, if a confounding variable decides to start a fight or gets too cozy with one of the other predictors or covariates (ie. collinearity/codependence) then it needs to be dealt with like any other predictor.

Key 2: Variable roles determine their location in the research question.

While figuring out which variables need to be included in your model, know that the variable roles described above will determine their location in your final research question. This will not only help you write out how the model is put together in your research description, but will help you with results reporting and determination of variable trimming if results indicate that variable loss would strengthen the model.

Outcome = Predictor1 + Predictor2 + Confounding + Covariate

DV = IV1 + IV2 + Confounding + Covariate

Key 3: Research question structure informs the analysis type (the next section).

By writing out the research question structure described in Key 2, you will be able to get a visual handle on your model which will inform the final decision as to what procedures will be used.

SAS PROCEDURES

The SAS procedures used in this section are very similar to data exploration. The difference is that instead of focusing on each variable by itself, you will be focusing on how the variables relate to each other. This will change the structure of your code slightly, and will add a few more steps into the exploration side of model preparation.

Commonly used procedures include FREQ, MEANS, UNIVARIATE, and CORR. It is important to familiarize yourself with these procedures and the different output and options they offer. It is in this step that you would pay particular attention to significance testing. You will want to see if there are significant relationships betweenyour variables of interest.

```
/* WHO Global Table Data */


PROC IMPORT OUT= WHOdata_raw

            DATAFILE= "D:\[Conference] Current Papers\Back to Basics\Resources - Data\WHO Global Table Data\WHO COVID-19 global table data January 21st 2021 at 8.01.21 PM.xlsx"

            DBMS=XLSX REPLACE; SHEET='WHO COVID-19 global table data '; GETNAMES=YES;

RUN;


PROC EXPORT

  DATA=WHOdata_raw

  DBMS=xlsx

  OUTFILE="D:\[Conference] Current Papers\Back to Basics\Resources - Data\WHO Global Table Data\COVID-19 WHO Data &sysdate..xlsx"

  REPLACE;

  SHEET='Final Dataset';

RUN;
```

```
data WHOdata;
	set WHOdata_raw;
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;
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;
```

During this stage, if you find necessity to adjust a variable to better fit your model, or clean up a bit of messiness, the DATA step can be employed. Most of the cleaning should have been completed in the prior step, but in the formal layout of a model, you may need to adjust variables in order to reduce error. This includes standardizing, binning, trimming outliers, or combining group categories. These adjustments are not only carried through in an easy-to-understand manner by SAS, but the existence of these adjustments within your code serve as archived documentation to how your data is being adjusted. This documentation will be of use when writing about your results or responding to an audit on your research.

There are plenty of other helpful procedures in SAS that are equipped to assist you during this step. Some popular ones are: CONTENTS, SORT, SQL, PRINT, and SAS Macro processing.

BEST PRACTICES

Practice 1: Document and implement findings from past research.

What has been done? Check the work of others for guidance on variable relationships and address these findings in your models. If you followed the scientific method, you should have

already completed a good portion of research in the third step. This is where you implement what you have found. Not only does the whole of your paper introduction address this step, but the results feed into your variable choices and the final model structure.

Practice 2: Variable couples counselling.

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. These relationships may be directly addressed in the assumptions of a particular model structure. The existence of unchecked relationships between predictors in a model is a direct violation of the common assumptions held by many of the most popular models, so will probably need to be addressed. Strongly correlated variables, especially if there are multiple variables with strong correlations, will need to be addressed as there are multiple ways to appropriately represent these relationships while minimizing their negative effect on the model.

Practice 3: Data structure incompatibility – mathematical theory.

Consider the differences between numeric and categorical data structure. These two data types have very unique interactions, especially if you consider how their sub-types interact with each other. Make sure to consider the limitations of mixing within-group data structures, Nominal & Ordinal, Multi-Level & Binary/Dichotomous, Interval & Ratio, or Discrete vs Continuous. Even the basic relationship between a dichotomous and 5-level categorical variable can result in boundary shrinkage, which negatively reflects a variable's impact in a model. The use and mathematical implications of using several different variable types needs to be explored. No matter the construct of your model, there are several ways to address miss-matching of variable types. You can employ procedures such as binning (for numeric to categorical conversion) and normalizing, or you can explore one of the several specialty designed macros that have been created to address the more touchy of variable interactions.

Practice 4: Always review your log.

SAS's log does a fantastic job of notifying you when it runs into an error or if the data is toeing the line of a procedure's underlying methodology. The log is not exhaustive and does not catch everything, but is a great first step in making sure everything ran the way it should. If you think you read this before, you did, this is how important it is. Be prepared to see it again.

STEP 4: MODEL DEPLOYMENT - MATCHING YOUR QUESTION TO A MODEL

Given the overlap between model building and deployment, this section builds the bridge between these two steps and helps smooth the journey from model to results.

KEY CONSIDERATIONS

Key 1: Check your model assumptions.

Make sure your model assumptions fit your question and data. Every model has its own set of assumptions and violation of these assumptions may lead to incorrect conclusions. Below are some commonly seen model assumptions.

Assumptions of Normality: Most parametric tests require that the assumption of normality be met. Normality means that the test is normally distributed (or bell-shaped) with mean 0, standard deviation 1, and a symmetric bell shaped curve.

Assumptions of Homogeneity of Variance: The assumption of homogeneity of variance states that the variance within each of the populations is equal.

Assumptions of Homogeneity of Variance-Covariance Matrices: The assumption for a multivariate approach states that the vector of the dependent variables follow a multivariate normal distribution, and the variance-covariance matrices are equal across the cells formed by the between-subjects effects.

Assumption of Linear Relationships: The assumption of linear relationships for linear regression states that the relationship between independent and dependent variables must be linear. The assumption of linear relationships for logistic regression states that the relationship between independent variables and their log odds must be linear.

Assumption of the Absence of Multicollinearity: Independent variables should not be highly correlated with each other.

Assumption of the Absence of Auto-Correlation: Residuals should be independent from each other.

Assumption of Randomization: The data collected must be a random sample from your population of interest. In order for a dataset to be a truly random sample, each subject must have an equal chance of being selected from the overall population.

Assumption of Large Sample Size: An adequately large sample size is required by some models, including logistic regression. Smaller sample sizes will under-power the procedure and the results will not be reliable.

Key 2: Mitigate violations.

If you found an assumption violation, not all is lost. It IS important to mitigate this violation. There are several different methods to do this, most of which involve some form of variable adjustment. Fatal violations (such as randomization) can not be mitigated through adjustment and will require a change in model choice.

SAS PROCEDURES

There are multiple SAS procedures for every assumption type. Some of the more specialized assumptions may require processes that span multiple steps. Check online documentation, as chances are, someone has had to address this violation before and has come up with a solution!

Below is a common way to address the assumption of multicollinearity. Through the REG procedure, you can call for options VIF, TOL, and COLLIN, which will give diagnostic estimates of the presence of multicollinearity.

```
/* ICPSR COVID Isolation on Healthcare Workers */


PROC IMPORT OUT= ICPSRdata_raw
            DATAFILE= "D:\[Conference] Current Papers\Back to Basics\Resources - Data\Open ICPSR\Copy of COVID Isolation on Sleep and Health in Healthcare Workers_April 29, 2020_08.17.xlsx"
            DBMS=XLSX REPLACE; SHEET='Adjusted'; GETNAMES=YES;
RUN;


data ICPSRdata;
      set ICPSRdata_raw;


      Q12a = Q12*1;
      Q13a = Q13*1;
      Q19a = Q19*1;
      Q20a = Q20*1;
      Q21a = Q21*1;
```

```
        Q22a = Q22*1;
run;


/* 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;
```

BEST PRACTICES

Practice 1: More than one way….

There are numerous routes to test an assumption. I recommend using multiple routes (if available) or narrowing in on the most appropriate one for your particular project (if more specialized).

Practice 2: Do not hesitate to switch models if needed.

If the assumption violations are too much for a particular model, then do not hesitate to switch to a different one. It is not your fault, the model's fault, or the data's fault, just take it as the opportunity to employ the best model for your particular question. This may mean stepping outside of your comfort zone, but you will have a better model for it.

Practice 3: Do not force a model.

This goes hand-in-hand with Practice 2. Do not try to force a model to happen. If you have to go through multiple variable adjustments just to fit a particular modeling type, then consider that that type might not be best for the question/data at hand and consider a different route. By adjusting a variable, you are still introducing an element of bias (as you are still adjusting the variable from its base form due to performance). Too many adjustments may lead you down a path not represented by your data. You must be able to interpret the end results in a meaningful way, and too many adjustments will certainly make that difficult.

Practice 4: Know that there is a high rollers club.

The more complex the analysis, generally the more numerous and complex the assumptions. Violations of these assumptions can be especially detrimental and may require significant steps to mitigate. Naïve implementation of these higher order analytic processes can also be very harmful. Do not hesitate to enlist theoretical help, as these analyses may have high pay out, but also come with great risk.

Practice 5: Null results do NOT mean model failure.

As stated in the introduction, you are more than likely seeking to reject the null hypothesis, the opposite of which would be to "fail to reject the null hypothesis". This is not a true failure, as the model and data did exactly what you asked them to do, they tested a relationship. If the relationship was null, or non-significant, then the model did its job and showed you that. These results are still meaningful results. Just maybe not the results you were hoping for!

Practice 6: Always review your log.

SAS's log does a fantastic job of notifying you when it runs into an error or if the data is toeing the line of a procedure's underlying methodology. The log is not exhaustive and does not catch everything, but is a great first step in making sure everything ran the way it should. If you think you read this before, you did, this is how important it is. I hope by now you can predict what the last best practice will be.

STEP 5: FEED THE LOOP - EVALUATE YOUR MODEL

By now, you have confidently run your model and have some results to work with. That's it right? Not really. You can still do one more thing to make sure your model is performing at its best. Next up: model fit and predictive power tests.

KEY CONSIDERATIONS

Key 1: After the model is run, you are not done!

Before giving your final report, you may want to check your model performance and validity of any new tests that have been incorporated into your research (if you conducted a prospective study using a new instrument). Can you think of any other ways to help strengthen this final report?

Key 2: Check model predictive power.

Measures of predictive power typically have values that fall between 0 and 1, with 0 indicating a complete lack of predictive power and 1 indicating a perfect predictive relationship. As a general rule, the higher the value, the better, but other than that there are rarely any fixed cut-off values that differentiate whether a model is acceptable or not.

Key 3: Check model fit measures.

Goodness-of-fit measures are formal tests of the null hypothesis that the fitted model is correct. These measures output a p-value which is used to decide whether or not the indicated model is a good fit. P-values are numbers between 0 and 1 with higher values indicating a better fit. Contrary to the traditional view of p-values, where one would specify a target alpha level (such as .05) and accept a model with a p-value below this value, goodness-of-fit test p-values that land below the specified alpha level would indicate that a model is not acceptable.

Key 4: Consider checking both predictive power and model fit, not just one.

It is important to note that goodness-of-fit measures and predictive power measures are testing two very different concepts. It should not be surprising for a model that has a very high R-square to also produce an unacceptable goodness-of-fit statistic. The opposite is also true, with the common existence of models with very low R-square and ideal goodness-of-fit scores. One way to look at these two concepts is like this: R-square scores test how much of the variation in response seen in the outcome variable can be explained by the proposed model, whereas goodness-of-fit scores do not tell you how well the outcome variable is predicted by the model, but rather if a specific model can do a better job at explaining the relationship between predictor and outcome than a previously proposed model. Through goodness-of-fit tests, the analyst can qualitatively compare different models while exploring the utilization of more complex concepts such as the addition of non-linearities, interactions, or changing the link function.

Key 5: If necessary, evaluate validity, reliability, and generalizability of data.

These are important concepts that will be questioned in a formal presentation of your research. Did you use a new test of procedure? How does this test or procedure measure up to the gold standard? Is your data and analysis able to be generalized to the larger population?

SAS PROCEDURES

Every modeling procedure has a way to test for predictive power and model fit, this is how important these concepts are. There are even procedures designed to manipulate and recommend a model based primarily on these measures!

For predictive power, some common procedures available as either default measurements or options include: Cox-Snell (LOGISTIC and REG), Tjur (LOGISTIC and TTEST), and many others.

As for model fit, there are many more procedures available. The use of these procedures is usually subject to the model type. For example: Pearson, Hosmer-Lemeshow, and Stukel are all options that can be explored in the LOGISTIC procedure aside from the defaults of AIC, BIC, and -2LogL. There are also numerous macros (such as the %goflogit macro) that have been created in order to run several tests at one time, making results comparisons even faster.

```
/* ICPSR COVID Isolation on Healthcare Workers */


PROC IMPORT OUT= ILVdata_raw
            DATAFILE= "D:\[Conference] Current Papers\Back to Basics\Resources - 
Data\Illinois Vaccine\Illinois Vaccine Data (2021.02.04).xlsx"
            DBMS=XLSX REPLACE; SHEET='Sheet1'; GETNAMES=YES;
RUN;


data ILVdata;
      set ILVdata_raw;


      Total_Reported_Inventory2 = Total_Reported_Inventory*1;
run;


proc contents data=ILVdata;
run;


proc print data=ILVdata;
run;


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;


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;
```

BEST PRACTICES

Practice 1: More than one way….

There are numerous routes to test for predictive power and model fit. I recommend using multiple routes (if available) or narrowing in on the most appropriate one for your particular project (if more specialized).

Practice 2: Do not hesitate to switch or restructure a model if needed.

If the predictive power or model fit reveal an unstable model, then do not hesitate to switch to a different one (if appropriate). It is not your fault, the model's fault, or the data's fault, just take it as the opportunity to employ the best model for your particular question. This may mean stepping outside of your comfort zone, but you will have a better model for it. It may also mean allowing the addition or removal of a variable you would have otherwise hoped to handle differently.

Practice 3: Do not force a model.

This goes hand-in-hand with Practice 2. Do not try to force a model to happen. If a model is weak or unstable, do not adjust your model simply to make it work. Accept it for what it is.

Practice 4: Null results do NOT mean model failure.

As stated in the introduction, you are more than likely seeking to reject the null hypothesis, the opposite of which would be to "fail to reject the null hypothesis". This is not a true failure, as the model and data did exactly what you asked them to do, they tested a relationship. If the relationship was null, or non-significant, then the model did its job and showed you that. These results are still meaningful results. Just maybe not the results you were hoping for!

Practice 5: Always review your log.

SAS's log does a fantastic job of notifying you when it runs into an error or if the data is toeing the line of a procedure's underlying methodology. The log is not exhaustive and does not catch everything, but is a great first step in making sure everything ran the way it should. If you think you read this before, you did, this is how important it is. I hope by now you can predict what the last best practice will be.

Best Practices: There is more than one way to test power and model fit – use them, do not hesitate to switch models, do not force a model, do not hesitate to appropriately restructure a model in poor health, null results do not mean model failure/incompatibility.

CONCLUSION

Carefully consider and implement the individual steps of the analytic process. From choosing/importing data, data exploration, data driven modeling, matching your question to your model, and model evaluation, each step has its place and can not be overlooked or completed halfway.

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.

REFERENCES AND RECOMMENDED READING

Given that this is an introductory paper, here is a list of great books and papers to help you on your journey to analytic exploration!

FURTHER READING

Cody, R. 2019. *A Gentle Introduction to Statistics Using SAS Studio.* Cary, NC : SAS Institute Inc.

Figard, S. 2019. *Introduction to Biostatistics with JMP.* Cary, NC : SAS Institute Inc.

Blum, J; Duggins, J. 2019. *Fundamentals of Programming in SAS: A Case Studies Approach.* Cary, NC : SAS Institute Inc.

Carver, R. 2019. *Practical Data Analysis with JMP.* Cary, NC : SAS Institute Inc.

Faries, D; Zhang, X; Kadziola, Z; Siebert, U; Kuehne, F; Obenchain, R; Haro, J. 2020. *Real World Health Care Data Analysis Causal Methods and Implementation Using.* Cary, NC : SAS Institute Inc.

Jansen, L. "SAS Conference Proceedings (1976 – present) … and more". Continuously updated. Available at https://lexjansen.com/.

Cathy O'Neil Books

RECOMMENDED FUN & INFORMATIVE BOOKS

O'Neil, C. 2013. *On Being a Data Skeptic*. O'Reilly Media. ISBN 1491947233.

O'Neil, C; Schutt, R. 2013. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media. ISBN 1449358659.

O'Neil, C. 2016. *Weapons of Math Destruction*. Crown. ISBN 0553418815

COVID DATASETS

CDC (2021). *COVID Data Tracker* [data file and codebook]. https://covid.cdc.gov/covid-data-tracker/#datatracker-home

Healthdata.gov (2021). *Assorted COVID Datasets* [data files and codebooks]. https://healthdata.gov/search/type/dataset?query=covid

WHO (2021). *Coronavirus Disease (COVID-19) Dashboard [*data files and codebooks]. https://covid19.who.int/?gclid=Cj0KCQiA1KiBBhCcARIsAPWqoSrOV1fyleLks1OsxtFmPiW9v7cxNJnQ3Gow5ZtFOY3BzICqZvjuyFwaArL4EALw_wcB

Kaggle, AI2, CZI, MSR, Georgetown, NIH, White House (2021). *COVID-19 Open Research Dataset Challenge (CORD-19)* [data files and codebooks]. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

NIH (2021). *Open-Access Data and Computational Resources to Address COVID-19* [data files and codebooks]. https://datascience.nih.gov/covid-19-open-access-resources

Illinois Department of Public Health (2021). *COVID-19 Vaccine Administration Data* [data files and information]. https://www.dph.illinois.gov/covid19/vaccinedata?county=Illinois – Example of Individual state-based repository

John's Hopkins (2021). *Finding Datasets for Secondary Analysis* [data files and codebooks]. https://browse.welch.jhmi.edu/datasets/Covid19

Open ICPSR (2021). *COVID-19 Data Repository: Data examining the impact of the novel coronavirus global pandemic* [data files and codebooks]. https://www.openicpsr.org/openicpsr/covid19

MIDAS (2021). *Online Portal for COVID-19 Modeling Research* [data files and codebooks]. https://midasnetwork.us/covid-19/

Google Cloud, Big Query (2021). *COVID-19 public datasets: our continued commitment to open, accessible data* [data files and codebooks]. https://cloud.google.com/blog/products/data-analytics/publicly-available-covid-19-data-for-analytics

ACKNOWLEDGMENTS

The author would like to thank all of the healthcare workers, researchers, epidemiologists and analysts who are working together to address the coronavirus pandemic. We would be in a far different position without you.

The author would like to thank all of the agencies who have made their coronavirus-related data accessible to the public. For those of us who have not had the honor to work on the frontlines, it has been a relief to be able to assist in our limited capacities. Making this data available has, hopefully, enabled us all to help out in some way.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Deanna Schreiber-Gregory, MS
Independent Consultant – Statistics, Data Management, Research Methods
Juxdapoze, LLC
d.n.schreibergregory@gmail.com or juxdapoze.consulting@gmail.com
www.juxdapoze.com