SAS® GLOBAL FORUM 2020 #SASGF

Paper 4747-2020

From Santiago to the Spirit Lake Nation: 30 **T**hings I **L**earned in 30 **Y**ears as a **S**tatistical **C**onsultant

AnnMaria De Mars, The Julia Group

ABSTRACT

All consultants face a few common problems; getting clients, getting the data, getting to know their data, fixing data problems, finally getting to the statistical analysis and communicating the results to an audience. After getting paid, communication is the most important part of the process. If your client doesn't understand your analysis, the findings aren't going to be applied and you're not going to get repeat business. Fortunately, SAS® procedures from statistical graphics to PROC FORMAT ease translation of results, leaving both you and the client happy. A major difference between mathematical statistics and statistical consulting is the messiness of the data. The ability to recognize and rectify data problems is a major factor in success as a consultant. Fortunately, again, SAS offers a wealth of possibilities for testing and improving data quality. Your clients count on you to have performed the right statistical procedures in the right way. SAS offers an enormous range of sophisticated procedures, but beware of the simple things that can trip you up.

INTRODUCTION

Statistical consulting is a broad field, from careers with a single client in a single industry, to generalists who cross time zones like most people cross town. All consultants face a few common problems; getting clients, getting the data, getting to know their data, fixing the #$@! data, finally getting to the statistical analysis and communicating the results to an audience. We'll cover clients first because if you don't have clients, you're not a consultant.

Analysis and communicating the results are the two most important parts of the process. Statistical analysis is to presentation like sin is to confession. If you don't do the one, you have nothing to say in the other. If your client doesn't understand your analysis, the findings are not going to be applied, and equally disappointing, you're probably not going to get repeat business. Fortunately, SAS procedures from statistical graphics to PROC FORMAT ease translation of results.

Once you have the data, it's just a matter of analysis, right? Let me answer after I pick myself up off the floor laughing. The major difference between mathematical statistics and applied statistics is the messiness of the data, and statistical consultants are applied statisticians. Common errors include overlooking out-of-range values, failing to reverse-code items, failing to recode missing data. For the issues that can be solved with SAS programming, macro code can automate common tasks. For errors that require thinking, you're on your own.

Issues in statistical analysis will depend on the statistical question, but the most common ones, covered in this presentation, include too much data (seriously, what were you thinking of doing with those 600 variables?), correlated data, missing data and too small of a sample size. Fortunately, SAS procedures has both to identify these problems and help to solve them.

These topics are covered below in order of importance. By "importance" I mean relevance to you staying in business.

THE BUSINESS SIDE OF THE CONSULTING BUSINESS

A CONSULTANT HAS PAYING CLIENTS.

Fun fact: the father of psychoanalysis, Sigmund Freud, only accepted paying clients (Crews, 2017). I'm telling you this because too many aspiring consultants think their biggest problem when they start out is going to be whether they should do propensity score matching using the nearest neighbor or caliper method. In an otherwise great article, "Aspects of statistical consulting not taught by academia", Kennett and Thyregod (2006) don't mention getting paid at all.

Here are the biggest problems you'll face:

- Getting your first clients
- Getting paid
- Getting your data into shape
- Communicating results to your clients.

Statistical analysis, covered at the end of this paper, will be the least of your problems.

Getting clients

The four major ways to get clients are; referrals, as part of a consulting company, through your online presence and through an organization. In general, founding a company allows more autonomy, corporations pay better, universities have better benefits and working conditions. Your mileage may vary.

Even if you do want to be self-employed, there are advantages to starting out at an established company. You'll need to learn the business side of running a consulting business like developing a proposal, contracts and deliverables. You can learn it on your own or you can learn it while being paid by a consulting company. As far as referrals, a common source is professors or former professors. Let them know you are interested in consulting work! If someone can't afford my fees or I am fully booked, I will refer potential clients to students, former students or other new professionals. Another source of referrals is previous employers. It's not uncommon to leave an organization and then be brought back as a consultant on specific projects. A third source is conferences like SAS Global Forum. I've had people I've never met provide referrals because they attended a talk I gave on factor analysis or logistic regression. I've also written a blog for a dozen years, AnnMaria's blog (2008-2020), that has brought in referrals. I mostly write for my own amusement and for documentation, rambling on about SAS, statistics or business and swearing a fair bit. Better examples of blogs that are more focused specifically on attracting business while providing quality content are The Analysis Factor (2008-2020) and rforstats (Muenchen, 2020).

There is a lot of talk about the importance of soft skills to succeed in a technical field but I've found that a lot of getting referrals boils down to two things; don't be a jerk and communicate well. I have specifics on that. Keep reading.

Getting Paid

If you don't want to be responsible for finding the business, staying with a university or consulting firm for your entire career is a viable choice. In that case, finding clients may be someone else's problem. However, it's still in your best interest to be able to place a reasonable price on your time. The two parts to computing a price are how many hours you expect a project to take and how much is reasonable per hour. Sen (2020) provides a good general guideline on estimating percent effort. He makes the point that the percent effort may vary depending on the level of the statistical consultant. A project that might require 80 hours from a senior statistician could take a more junior professional 160 hours. The hours estimated also depend on what you can reasonably expect from the client. For example, whether the client can write up the method and results section without assistance once analysis is completed versus expecting the consultant to write those sections is going to make a difference in hours required.

What's a fair hourly rate?

> *Ask yourself, if I had twice as many grants/ contracts as I could do and I was paying someone to do this work, what would I be willing to pay?*

For a more scientific method, check the American Statistical Association survey for the median salary based on your degree and years of experience. Their latest report for academic salaries was completed in 2019 (Ange et al. 2019). The survey results for business, government and industry are available for 2015 salaries (Hall and George, 2017). If your rates are significantly higher than the average for your qualifications, unless there is some reason to justify that, you may find yourself priced out of the market.

Pro tip: Offering a discounted rate for multi-year contracts can provide some stability to independent consultants by insuring a minimum base salary.

Contracts and invoices

> *A verbal agreement isn't worth the paper it's printed on.*

You need a signed contract. The level of detail is going to depend on the amount of work. At a minimum, include how much you'll get paid, whether the client pays for travel or other expenses, the specific tasks, deadlines for each task and when payment will be due. That last part is often overlooked by new consultants. If you have a three-month contract, rather than waiting three months to get paid, break it down into three deliverables with payment due when each is received and approved by the client.

The sooner you send invoices, the sooner you get paid.

Take twenty minutes to find out the process of paying your invoice. If it needs to be approved by the principal investigator of a grant, then a purchase order is created by the administrative assistant, copy the assistant on the emailed invoice. Ask the assistant if there is a particular format or requirements of the invoice. Call the client's accounts payable department if you're not sure.

The Key to Success as a Statistical Consultant: Don't be a Jerk and Don't Be Too Clever

One reason a lot of consultants go bankrupt or have to find another line of work is they do think they are smarter than their clients. This manifests itself in a lot of ways. Being a jerk includes being "too important" to talk to support staff to find out how to make it easier for them to pay you, having unrealistic salary expectations, misuse of software licenses and overcharging for work done. If your former employer, Big Company X was charging clients $325 an hour for your time, it's not necessarily a fair price for you as an independent consultant unless you can provide the same level of service, including e.g., HIPAA certified servers, 24-hour response time for support requests.

I've met a lot of people over the years who charged much more than me and bragged to me about it. In the long run, though, I made more money. Charge a fair rate and when you have more work than you can do, hire employees, raise your rates and/or be more selective in the projects you choose. Maybe you really are in the top 1% of statistical consultants so that justifies charging top rates. However, if your client simply needs a repeated measures ANOVA run and interpreted correctly, they probably don't need to pay for someone who is having statistics named after them.

Don't Be Too Clever

Don't Charge for More than the Client Needs

I wrote, tested and documented a macro for data quality, discussed below. It took me two hours. In the long run, the macro saved me some time but it wouldn't be right to charge that to the first client for whom I used this macro. On the other hand, if it was a long term contract and I was going to be doing similar analyses many times, maybe it would.

You're not doing analyses to impress your former professors, classmates or colleagues at SAS Global Forum. If an ANOVA will do, don't do a structural equation model. If only 2% of the data are missing, you don't need PROC MI for multiple imputation.

Don't Steal Software

Misusing software licenses falls under the business side because it can get you sued. If you have a SAS academic license, for example, you are free to use that for research and teaching. If you do work for a client, in their facility, on their computer, or by logging into their computer and they have a site license for SAS, you're fine because the client is paying SAS for that license and you are using it as their contractor. If you have an academic license through your university and you use it for consulting for which you charge people, you owe the software company money.

THE FIVE BASICS OF CONSULTING SUCCESS

Successful consultants have five categories of skills; communication, testing, statistics, programming and generalist.

COMMUNICATION

Of all the qualities necessary to be a successful statistical consultant, none is more important than communication, even if that communication is only with your future self. All five skill sets are necessary to some extent, but a terrific communicator with mediocre statistical analysis skills will get more business than a stellar statistician that can't communicate.

Documentation

Communication includes documentation, both in your code and internal documents such as codebooks or an internal wiki. It includes letting clients know what you're going to do, what it's going to cost, what that cost includes, what were your results and what those results mean. If you're good at communicating with clients, colleagues and your future self, you're half-way to success.

An example of the critical nature of communication can be found in the following retraction:

> *The identified programming error was in a file used for preparation of the analytic data sets for statistical analysis and occurred while the variable referring to the study "arm" (i.e., group) assignment was recoded. The purpose of the recoding was to change the randomization assignment variable format of "1, 2" to a binary format of*

> *"0, 1." However, the assignment was made incorrectly and resulted in a reversed coding of the study groups." Aboumatar and Wise (2019, p. 1417)*

Because of this incorrect coding, the reported results were the exact opposite of what actually occurred.

Document coding!

Here is an example from a current research project where the CES-D depression scale was used, which requires several items to be reverse-coded before scoring.

In the HTML file where the user enters data that's written to the database there is this comment:

```
    <h5>I felt that I was just as good as other kids.</h5>
    <! –– This is reverse-coded. Don't you dare change it. ––>
<div class="row mb-3">
    <button id="cesd4_1" data-src="3" class="cesd4 btn btn-light
shadow-box col-5 my-3 mx-auto">Not at all</button>
```

In the original file to read in the data to SAS, there is a comment:

```
*** NOTE: CESD IS ALREADY REVERSE-CODED. DOES NOT NEED CODING!;
FILENAME REFFILE2
    '/home/directory3/examples/cesd.xlsx';
```

In the internal wiki, there is this note:

TABLES IN ACME PROJECT DATABASE

CESD - Center for Epidemiologic Studies Depression Scale - NOTE: The data are reverse coded at data entry. There is no need to reverse code these. There are 25 columns in this table; ID, username, session number, questions 1 through 20 of the CESD scale, the CESD total which is the sum of the 20 questions, named item21 and a time stamp.

Document everything! Document how are items coded, how subscales or totals are computed.

This may seem like overkill, but how many retractions could be prevented by this level of documentation? If you are a consultant, it's probable that at some point someone else will be looking at these data, or that you may be called back a year later to do a longitudinal analysis. Your colleagues and future you will thank you.

Communicating with graphs

If you're a statistician, when you look at a table for a repeated measures Analysis of Variance, the F-statistic, R-square and p-value for the time by treatment interaction are self-evident. In logistic regression, the interpretation of the confidence interval of the odds ratio is clear to you. However, your clients are NOT statisticians. If they were, they wouldn't need you.

Consider this code, using the dataset from Kaiser Permanente on the oldest old (Haan et al., 2011), which produces 8 pages of output.

```
        PROC LOGISTIC DATA=old PLOTS= ALL;
           CLASS nhome sex  ;
            MODEL dthflag =  agein71 nhome sex;
```

Let's explain the entire analysis with one graph. This graph shows the predicted probability of being alive. As age increases, from 65 to 95 the probability of still being alive at the end of the study decreased. The blue line represents women who were not in nursing homes. At every age, they are the least likely to die. The red line is men not in nursing homes. The green line is women in nursing homes, the brown line is men in nursing homes. A 65-year-old man in a nursing home has the same probability of dying as an 80-year-old woman who is not in a nursing home. Although both gender and nursing home residence are significant, you can see that the impact of nursing home status is greater than gender just by looking at the difference between the lines.

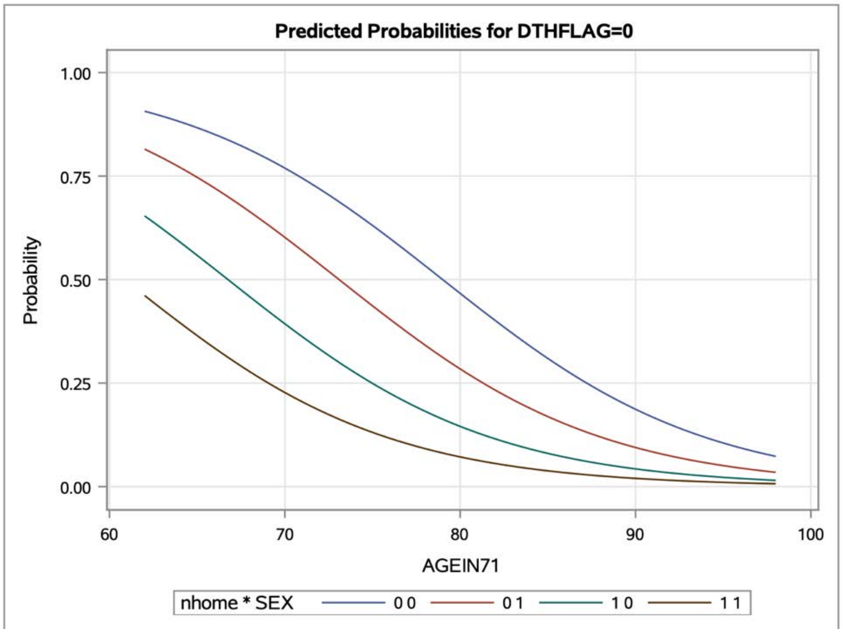


Figure 1 Predicted Probabilities from PROC LOGISTIC

Graphs are great but know your audience

With most clients, I would use the predicted probabilities plot in preference to the odds ratio plots if I had both continuous and binary variables. I do this because with mixed variable types, I would be explaining that age is not necessarily unimportant but rather that the odds ratio represents the change in the odds for a change in the value of a predictor, X. A change from 65 to 66 years is not equivalent to a change from 0 (= not in a nursing home) to 1(=in a home). To many clients, this is not self-evident nor nearly as easy to understand as the graph above.

I would not usually use Figure 2 below in a presentation to a client, but I would use Figure 3 if, instead of age as a variable I had happened to have the data as two groups, subjects over 75 and those 75 and under.

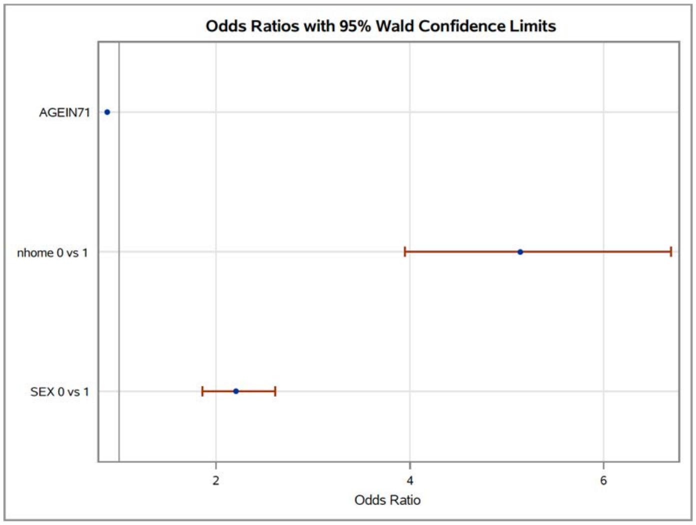


Figure 2 Odds ratio plot mixing continuous and binary variables

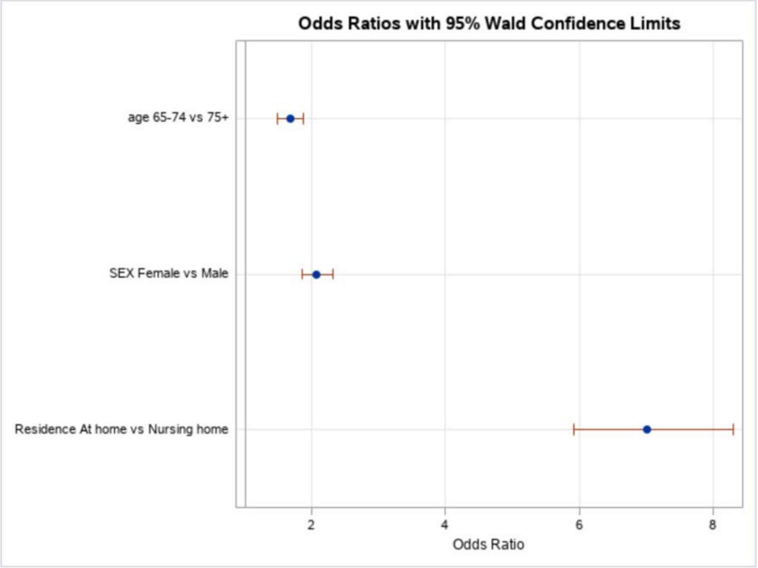


Figure 3 Odds ratio plot with binary variables only

For some clients, it may be easy to explain that the coefficient for a continuous variable is the expected change in the log odds for a one-unit increase in X. A difference of one unit from age 66 to age 67 doesn't have the same meaning as a difference from 0= female to 1 = male. Now, if you are a statistician, you are probably nodding your head and we could work through to exponentiating both sides of an equation to show how the odds ratio changes for one-unit increase of a continuous variable. That is not the case for most clients.

On the other hand, it's very easy to explain Figure 3. The first vertical line is at 1. That would mean the odds are equal. If the odds are the same, the odds for one group, say females, divided by the odds for the other = 1. You can see that all three variables are to the right of that line. The blue dot is the point estimate of the odds ratio and those red lines are the confidence intervals. If that red line crosses the line at 1 that means that 1 is within the 95% confidence interval. You can see that's not the case for any of these variables. Not only are all significantly different from 1 but two of the variables, age and gender have an odds ratio of around 2 and the third variable, living in a nursing home has an odds ratio above 6.

These are just a few examples of the huge number of options for effective use of plots. Others that lend themselves to easy communication of results include diagnostic plots in regression and ANOVA, plots of interaction effects, distribution plots in t-tests. Getting to know ODS Statistical Graphics is time well-spent.

Communication in reporting

Being able to communicate technical information to a non-technical audience is a valuable skill. Given the description above, the average person can tell that being in a nursing home has a greater predictive value for death than does age or gender.

This isn't 'dumbing down' a report for clients. Obviously, clients are smart enough to be in charge of a budget to pay statistical consultants. They probably don't, however, have the interpretation of a standardized versus unstandardized beta weight memorized. The higher up the organizational chart your results are read, the less time the reader is going to spend on it and the less likely they'll be a statistician. Include the global null hypothesis tests, AIC and Wald chi-square in an appendix. It will be available for the specialist who reads your report in detail while everyone else will be able to quickly comprehend the information they need from the main body.

When data are binary or categorical, use PROC FORMAT

Another reason that Figure 3 is preferable to Figure 2 is that it's clear what values are being compared. Three months from now, odds are no one is going to remember whether 0 was female or male. The format procedure handles this nicely.

```
PROC FORMAT ;
  VALUE gender
  0 = "Female"
  1 = "Male" ;
  VALUE nhome
   0 = "At home"
   1 = "Nursing home";
  VALUE old
    0 = "65-74"
    1 = "75+" ;
```

Does "nhome" = 1 mean the subject is in a nursing home or does it mean "in the home"? Solve the issue by pairing the PROC FORMAT with a RENAME statement in the DATA step.

```
RENAME nhome = Residence ;
```

These statements make the results much clearer at a glance not only to the client but also the statistical consultant. Don't underestimate the importance of these simple changes. While, e.g. using a Scheffé vs Tukey post hoc test in an analysis might alter your p-values slightly, the decision isn't going to completely alter your results in the same way as confusing the coding of your variables.

TEST EVERYTHING

For anyone writing software, testing is important. However much you think you need to test your code, you are wrong. The answer is, "More." For independent consultants this is even more of an issue because you may be the only person reviewing this code before it goes into production or the results are used to develop policy.

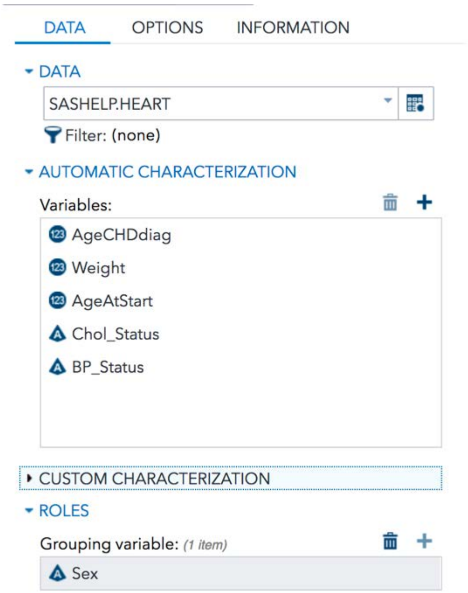
Step 0: Coding and Documentation

Documentation is the pre-alpha level of testing. Commenting or otherwise documenting your code often leads to quality control moments from “Oh, !@$& , I didn’t reverse code those variables!” to implementing macros or do-loops to reduce duplicate code.

Step 1a: Easy Quality Control with the Characterize Data Task

For a quick and easy first look at a small number of variables, the Characterize Data task is useful. It’s found in SAS Studio under Tasks > Data. Simply select the data set, then add the variables of interested by clicking the + sign. Statistics can be requested overall or by a grouping variable. The Characterize Data task automatically generates the code to produce frequency distributions with plots for all character variables using PROC FREQ. For numeric variables, N. number missing, minimum, mean, median, maximum and standard deviation are computed using PROC MEANS and histograms are produced with PROC UNIVARIATE.

Quick tip: SAS will compute statistics based on type so if race is in your dataset as numeric coded 1=African-American, 2=Asian-American, etc. the Characterize Data task will produce means and standard deviations for race. Be smarter than your computer and delete these statistics before you include results in any report.



DATA

SASHELP.HEART

Filter: (none)

AUTOMATIC CHARACTERIZATION

Variables:

- AgeCHDdiag
- Weight
- AgeAtStart
- Chol_Status
- BP_Status

CUSTOM CHARACTERIZATION

ROLES

Grouping variable: *(1 item)*

- Sex

Figure 4 Characterize Data Task Window

Step 1b: Quality Control with Descriptive Statistics

If you don’t want to click 512 times for all of your variables, there are the PROC MEANS, PROC SUMMARY and PROC UNIVARIATE tasks. For each of these, if no VAR statement is included, descriptive statistics are computed for all numeric variables. In addition to the mean, minimum, maximum, N, number missing and standard deviation produced by those descriptive procedures, PROC FREQ is useful and underutilized for checking the number of levels. For example, I have 622 records in this particular data set. I’d like to know if there are any duplicate IDs. An easy way to do this is:

```
      PROC FREQ DATA=mydata2.example1 NLEVELS ;
          TABLES userid / MISSING;
```

Don’t omit the missing option! The SAS log shows

NOTE: There were 622 observations read from the data set MYDATA2.EXAMPLE1.

The output from my PROC FREQ shows

Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
userid	userid	514	1	513

Table 1 Output from PROC FREQ with NLEVELS

I can see from Table 1 that I have 514 levels, including ‘missing’. With 622 observations, this appears I have quite a few duplicate IDs.

However, looking at the missing level in Table 2, it’s clear that nearly all of those duplicate IDs are simply missing. Even deleting those 85 duplicates from the 622 records, that still leaves 537 records for the other 513 levels, so there are some duplicate IDs.

userid

userid	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	86	13.83	86	13.83
1234567	1	0.16	87	13.99
16mhallow	1	0.16	88	14.15

Table 2: Partial output from PROC FREQ with MISSING option

Test for Missing and Invalid Data

Checking a dataset with hundreds of variables is a recipe for human error. A few lines of SAS code automate the task. In the steps below, first, a PROC FREQ outputs duplicate ID numbers to a data set and the PROC PRINT step prints the first 10.

```
      TITLE "Duplicate ID Numbers" ;
      PROC FREQ data =  lib.adult2009 NOPRINT ;
            TABLES puf_id / out = adult_freq (WHERE = ( COUNT > 1 )) ;
      PROC PRINT DATA = adult_freq (OBS = 10 ) ;
```

It turns out there are not any, a fact that can be seen from the SAS log.

```
NOTE: There were 47614 observations read from the data set LIB.ADULT2009.
 NOTE: The data set WORK.ADULT_FREQ has 0 observations and 3 variables.
```

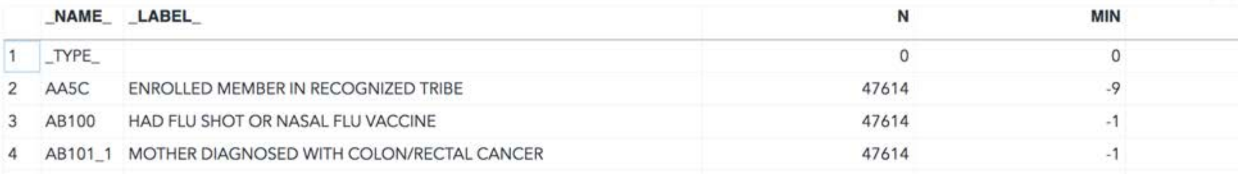
PROC MEANS produces mean, minimum, N and standard deviation for all numeric variables in the data set.

```
      PROC MEANS DATA = lib.adult2009 MEAN MIN N STD ;
      OUTPUT OUT = adult_stats ;
```

Next, the dataset is transposed

```
PROC TRANSPOSE DATA = adult_stats OUT = adult_stats_trans ;
      ID _STAT_  ;
```

The transposed dataset has a row for each variable and a column for each statistic



	NAME	_LABEL_	N	MIN
1	_TYPE_		0	0
2	AA5C	ENROLLED MEMBER IN RECOGNIZED TRIBE	47614	-9
3	AB100	HAD FLU SHOT OR NASAL FLU VACCINE	47614	-1
4	AB101_1	MOTHER DIAGNOSED WITH COLON/RECTAL CANCER	47614	-1

Figure 5 Dataset created by PROC TRANSPOSE

With some simple arithmetic, all variables below a certain threshold for missing, with a 0 standard deviation or a minimum less than 0 can be computed.

```
DATA adult_chk ;
      SET adult_stats_trans ;
      pctmiss = 1 - (n/47614) ;
      IF min < 0 THEN neg_min = 1 ;
            ELSE neg_min = 0 ;
      IF std = 0 THEN constant = 1 ;
            ELSE constant = 0 ;
      IF (pctmiss > .05 OR neg_min = 1 OR constant = 1) THEN OUTPUT ;
TITLE "Deviant variables to check " ;
PROC PRINT DATA = adult_chk ;
```

This code results in ALL of the variables being selected! Looking at the results reveals a problem. The minimum is negative for every variable. For many variables in this data set, responses were coded as 1= yes, 2=no and negative numbers were used for various non-responses such as refused, didn't know or not applicable. All of these variables have a type of 'numeric'. Without a detailed knowledge of the data, analyses using these variables as numeric predictors are going to significantly impacted by these outliers of -8 or -9 on a binary scale!

Obs	_NAME_	_LABEL_	N	MIN	MAX	MEAN	STD	pctmiss	neg_min	constant
1	_TYPE_		0	0	0	0.00	0.00	1	0	1
2	AA5C	ENROLLED MEMBER IN RECOGNIZED TRIBE	47614	-9	2	-0.94	0.56	0	1	0
3	AB100	HAD FLU SHOT OR NASAL FLU VACCINE	47614	-1	3	-0.07	1.04	0	1	0
4	AB101_1	MOTHER DIAGNOSED WITH COLON/RECTAL CANCER	47614	-1	2	-0.83	0.65	0	1	0

Table 3: Output of Suspect Variables from Data Quality Analysis

In the example above, items are checked for a minimum less than zero. The same code can be modified for checking items out of range on a scale, e.g. for a minimum less than 1 or greater than 4.

A macro for checking data quality

All of the steps above can be combined into the following macro

```
%SYSMSTORECLEAR;
OPTIONS MSTORED SASMSTORE=maclib;
LIBNAME maclib "directory to store macro" ;
LIBNAME lib "data directory" ;
%MACRO dataqual(dsn,idvar,obsnum) / STORE ;
Title "Duplicate ID Numbers" ;
PROC FREQ DATA =  lib.&dsn noprint ;
  TABLES &idvar / out = &dsn._freq (WHERE = ( COUNT > 1 )) ;
```

```
      FORMAT &idvar  ;
      PROC PRINT DATA = &dsn._freq (obs = 10 ) ;
      PROC MEANS DATA = lib.&dsn MEAN MIN N STD ;
        OUTPUT OUT = &dsn._stats ;
      PROC TRANSPOSE DATA = &dsn._stats OUT = &dsn._stats_trans ;
         ID _STAT_ ;
      DATA &dsn._chk ;
         SET &dsn._stats_trans ;
          pctmiss = 1 - (n/&obsnum) ;
          IF min < 0 THEN neg_min = 1 ;
              ELSE neg_min = 0 ;
          IF std = 0 THEN constant = 1 ;
             ELSE constant = 0 ;
          IF (pctmiss > .05 or neg_min = 1 or constant = 1) THEN OUTPUT ;
      TITLE "Deviant variables to check " ;
      PROC PRINT DATA = &dsn._chk ;
      TITLE "First 10 observations with ALL of the variables " ;
      PROC PRINT DATA = lib.&dsn (obs= 10)  ;
      RUN ;
      %MEND dataqual ;
      %dataqual(all_answers18,username,30848) ; * Call macro ;
```

This macro also prints out the first 10 observations from the data set. Egregious errors, for example, not realizing that data are missing for >90% of the subjects for many variables, can often be spotted by simply looking at the first few observations. For line by line breakdown of this macro, see DeMars (2020b).

Arrays to Fix Multiple Variables with Missing Data

Fixing missing data across hundreds of variables is very simple. It’s also often useful to identify subjects who have missing data, both to test for if data are missing at random and also to consider deleting subjects from an analysis if too many items need to be imputed. This code addresses both needs.

```
DATA adult ;
      SET lib.adult2009 ;
      ARRAY fixdata {*} _NUMERIC_ ;
      missdata= 0;
      DO i = 1 to DIM(fixdata);
            IF fixdata{i} < 0 THEN DO ;
                  missdata = missdata + 1;
                  fixdata{i} = . ;
            END;
      END;
```

Using the _numeric_ keyword creates an array of all of the numeric variables. The * will give the array a dimension of however many numeric variables are in the data set. This is a timesaver when a dataset has hundreds or thousands of variables and a mix of character and numeric variables.

The DIM function returns the dimension of the array. So, for from i = 1 to the dimension of the array, that is, for every numeric variable, if the value is less than 1 the missdata variable will be incremented by 1 then the value of the variable will be set to missing.

NOTE: The example above assumes that there are no variables where a negative number would be a valid score. If all items are scored on a scale of 1 to N, this is a valid assumption. In some cases, such as business income, a negative would be a valid value and should not be set to missing. Know your data!

Finding Data Errors with PROC CORR

While there is no guaranteed method to finding all data entry or coding errors, PROC CORR can be used to compute Cronbach's alpha using the code below.

```
PROC CORR DATA=example ALPHA ;
      VAR cesd1-cesd20 ;
```

There has been discussion of the limited usefulness and misinterpretation of Cronbach's alpha (Sijtsma, 2009) but one clear use is in identifying variables that were incorrectly coded. In this example, the variables were reverse-coded correctly but the first item read in was <u>not</u> actually the first item from the Center for Epidemiologic Studies depression scale but, in fact, the session number. Making this error even more difficult to identify, the session number ranged from 0 to 3, the same as the items on the scale. Further, the overall alpha is reasonably high.

Cronbach Coefficient Alpha

Variables	Alpha
Raw	0.911521
Standardized	0.912513

Table 4: Coefficient alpha results

The first clue something is not correct is the correlation of item cesd1 with the total, shown in Table 5 below, which is near zero and negative.

Cronbach Coefficient Alpha with Deleted Variable

Deleted Variable	Raw Variables		Standardized Variables		Label
	Correlation with Total	Alpha	Correlation with Total	Alpha	
cesd1	-.068726	0.922609	-.065748	0.922617	cesd1
cesd2	0.578660	0.906795	0.581331	0.907767	cesd2
cesd3	0.655730	0.905048	0.655204	0.905971	cesd3
cesd4	0.350876	0.912315	0.349347	0.913270	cesd4

Table 5: Output from PROC CORR with ALPHA option, Correlation with Total

Inter-item correlations for variable cesd1 are also all near-zero.

	cesd1	cesd2	cesd3	cesd4	cesd5	cesd6	cesd7	cesd8	cesd9
cesd1	1.00000	0.02558	0.12683	-0.11494	-0.03002	-0.13045	-0.09012	-0.19841	-0.06643
		0.8151	0.2446	0.2920	0.7838	0.2313	0.4092	0.0671	0.5434
cesd2	0.02558	1.00000	0.37762	0.20329	0.47549	0.33768	0.46842	0.16850	0.31886
	0.8151		0.0003	0.0605	<.0001	0.0015	<.0001	0.1209	0.0028

Table 6: Inter-item correlations

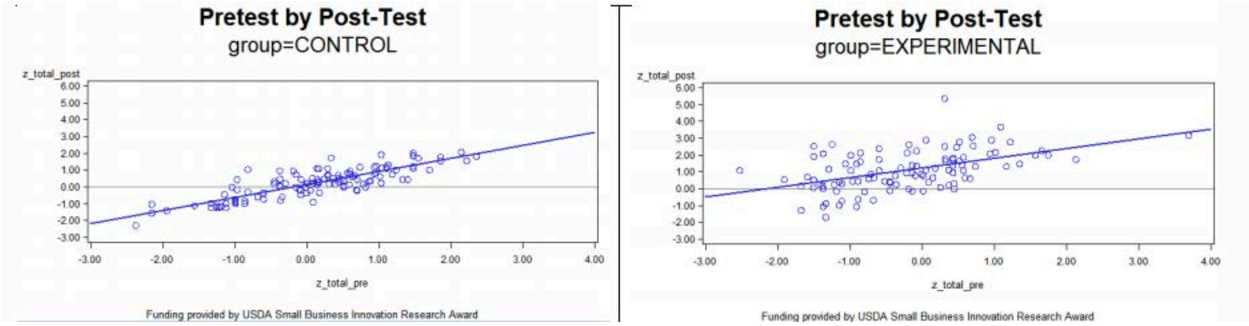
In this example, the item was simply read in incorrectly and did not belong in the scale. In cases when the item <u>does</u> belong in the scale but you fail to reverse-code only one item, a significant, negative correlation with other items in the scale and the total would be expected. If numerous items are incorrectly coded, a much lower than expected alpha would be observed, most items in the scale would have low item-total correlations and a mix of positive and negative inter-item correlations would be observed.

Finding Data Errors with PROC SGPLOT: Before you do repeated measures analysis, do this

```
GOPTIONS HBY = 2;
PROC SORT DATA = example ;
      BY group ;
PROC GPLOT DATA=example UNIFORM;
PLOT z_total_post * z_total_pre / VREF=0 ;
BY group;
```

Three points worth noting:

1. GOPTIONS HBY = 2 The default font size has the by-group title very small relative to the first title line. This option increases the font-size focus on the difference between the control and experimental group. Communication is always important.
2. PROC GPLOT DATA = datasetname UNIFORM - The UNIFORM option overrides the default which is to use the values to set the minimum and maximum of the axes. To facilitate comparison, the experimental and control groups have the same axes.
3. VREF = 0 - Draws a vertical line at the specified point, in this case 0 which, since these are z-scores, is the mean at the pretest, before the intervention. This reference line allows comparison, at a glance, to the proportions above and below the post-test mean.



Pretest by Post-Test

group=CONTROL

Funding provided by USDA Small Business Innovation Research Award

Pretest by Post-Test

group=EXPERIMENTAL

Funding provided by USDA Small Business Innovation Research Award

Figure 6 Plots of Pre- and Post-Test by Group

All of these questions can be answered easily looking at this graph:

- Did more of the experimental group increase their scores than the control group? Clearly, yes, as one would expect if the treatment was effective.
- Are there outliers that could impact the results? One person was nearly four standard deviations from the mean on the pretest. Another person was five standard deviations above the (pretest) mean on the post-test.
- Could any significant difference in favor of the experimental group be due to these outliers? Probably not. The control group is about equally above and below the mean at post-test while on the right almost all of the experimental group is above the mean. Still, it would be advisable to run, and report, these results with and without the outliers.
- Is there more scatter around the regression line for the experimental group than the control group?

The last point is particularly worth noting. In this study of work place training, the control group, which had no training, would be expected to score very similarly on the second occasion since subjects simply took the test twice. In the experimental group, more scatter would be expected as some people will benefit more from the training than the others either due to more attention in training or because they entered the training with a lower level of knowledge. If the reverse pattern is observed, or there are outliers that increased substantially, a discussion on coding with the data entry staff is warranted as well as investigation of the outliers.

STATISTICS FOR STATISTICAL CONSULTANTS

One of the qualities Kennett and Thyregood identify that a statistical consultant should have is "the confidence to use as simple a procedure as will get the job done". In addition to the procedures mentioned above for testing data quality, there are a few statistical procedures that fit this requirement. These are what I have found to be the "work horses" in terms of meeting the needs of the largest number of clients. Your mileage may vary.

Factor Analysis for When You Have Too Much Data

Factor analysis is a useful start for developing scales from the 50 items a client administered to 300 people. Do not do 50 chi-square analyses or you will burn in that corner of statistical hell reserved for people with outrageous Type I error.

Exploratory does not mean thoughtless!

```
*** Please do not do this BAD EXAMPLE 1;
      PROC FACTOR DATA = example ;
      RUN;
      ** Or this BAD EXAMPLE 2 ;
      PROC FACTOR DATA = example ;
          Var firstvar -- lastvar  ;
```

Unless you really, really know that every numeric variable in your data set including ID number can reasonably be expected to load on a factor, don't use bad example number one. Similarly, unless you have some rationale for expecting all of the variables from the first to the last in your list to correlate with some underlying factor, do not do bad example number two. Take this advice from someone who every semester asks students,

"This variable for race, coded 1= African-American, 2= Asian-American, etc. loads on the fifth factor. How do you interpret that?"

Factor analysis with PROC FACTOR can also be useful for testing validity of assumptions about data, e.g. that the first ten items your client wrote really do measure pain tolerance and the next ten measure anxiety. PROC CALIS (SAS Institute, 2013) is also useful and not overly complicated for confirmatory factor analysis.

The General Linear Model

Two of the three most common requirements clients will have are to predict a certain continuous outcome from multiple predictor variables. If the errors are normally distributed, as well as meeting other assumptions, multiple regression can be used.

```
      PROC REG DATA = example ;
            MODEL weight_lost = minutes_used education / STB;
```

While PROC REG and PROC GLM in SAS will give the same results in terms of model significance, F-value, etc. an advantage offered by PROC REG is the standardized beta weight option which enables direct comparison of the relative impact of predictors.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	26.24714	14.84446	1.77	0.0777	0
Minutes_used	Minutes App Used	1	-0.00041202	0.00069808	-0.59	0.5554	-0.02748
education	education	1	1.93815	0.37987	5.10	<.0001	0.23758

Table 7: PROC REG output with standardized beta weights

If one or more variables are categorical, PROC GLM can be used, with a CLASS statement identifying the categorical variables. Categorical variables can also be dummy-coded for 1=African-American , 0= Not African-American etc.

Of course, when a variable has N possible values, it should be coded as N-1 dummy variables. On occasion, the error may be made of using N dummy variables.

```
        ** BAD EXAMPLE 3, DON'T DO THIS! ;
        PROC REG DATA = example ;
                MODEL weight_lost = male female education;
```

It is nice to know that whether PROC GLM is used or PROC REG, the SAS output will include an error message in these cases. The error messages vary slightly and in my opinion the message from PROC REG is the clearest, as shown below.

Note:Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note:The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

female =	Intercept - male

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	19.53484	4.37565	4.46	<.0001
male		B	0.44615	1.24903	0.36	0.7211
female		0	0	.	.	.
education	education	1	1.78556	0.36732	4.86	<.0001

Table 7: PROC REG output with N dummy variables instead of N-1

Another data problem is when variables are not linear combinations but still highly correlated, for example, using both the year of admittance to a program and the number of days of attendance. To detect this problem use the VIF option for the variance inflation factor.

```
      PROC REG DATA = example ;
            MODEL earnings  = gender education admit_year days / VIF;
```

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	5396.83860	4368.74428	1.24	0.2174	0
Gender		1	0.45959	1.28963	0.36	0.7217	1.00205
education	education	1	1.89851	0.38134	4.98	<.0001	1.00779
admit_year	Year admitted	1	-2.74032	2.22909	-1.23	0.2196	76.95173
days	Days attended	1	0.00706	0.00612	1.15	0.2495	76.90668

Table 8: PROC REG output with variance inflation factor

A variance inflation factor over 10 indicates an issue with multicollinearity. For more on multicollinearity, how to detect and address it, see Schreiber-Gregory (2017).

Repeated Measures

The two gross violations of the assumption of uncorrelated data mentioned previously were due to errors but there is a common case where correlated data is expected and that is the case of repeated measures. The most common of these is a pre-post-test design. For repeated measures, SAS offers a lot of options, but the three most common, from simplest to slightly more complex would be PROC TTEST, PROC GLM and PROC MIXED. The paired ttest would be the least desirable design due to lack of a control group. PROC MIXED and PROC GLM generally produce identical results in terms of significance. I would argue, very strenuously, that the data issues discussed above generally have a far greater impact on the correctness of results than does a modest violation of the assumption of sphericity.

CATEGORICAL DATA ANALYSIS

Logistic regression pitfalls in SAS

The third common requirement clients have is to predict a categorical outcome from multiple predictor variables and the usual procedure is logistic regression. We already discussed this under graphs but there are two more points that deserve mention. In a binary logistic regression, by default, SAS predicts the lower value. Thus, if your dependent variable is dthflag with a value of 0 = alive and 1 = dead, SAS predicts survival, that is dthflag = 0. To change the value predicted, use the DESCENDING option.

```
      PROC LOGISTIC DATA =mydata.example  DESCENDING ;
```

In the SAS output, directly underneath the response profile table will be a statement of the probability being modeled, such as:

```
                Probability modeled is DTHFLAG='1'.
```

It is in tiny font, but still, this is an easy one to catch. Not as obvious, perhaps, is the fact that if a dependent variable has more than two categories the default is ordinal logistic regression. So, if you are predicting marital status with three categories for 1=married , 2 = divorced and 3= never married, these are going to be predicted using a cumulative logit model. To use a multinomial logistic regression, use the link=glogit option as shown below.

```
      PROC LOGISTIC DATA=mydata.adult09;
         CLASS educ income ;
         MODEL marit = educ bmi_p income / LINK=GLOGIT ;
```

Small data, No Problem

Maximum likelihood methods assume a relatively large sample size, which is not always available. A common situation in my experience has been when physicians in a single clinic are assessing the impact of a specific variable on mortality. Advice to have a few more patients die to increase sample size is not well received. Another recent need was early in my own research comparing barriers to implementing new technology in urban and rural school districts. The first year only 17 districts in the study area attempted major changes in technology. In these situations, chi-square would not be an appropriate test due to expected cell counts less than 5 in all of the cells. A Fisher's exact test would be appropriate and is requested as follows.

```
      PROC FREQ DATA = install ;
            TABLES rural*install / CHISQ ;
```

Yes, that is just a PROC FREQ with a chi-square. If you have a 2 x 2 table, SAS automatically computes the Fisher exact test, as well as several other tests. See, not all defaults in categorical data analysis are problematic.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.0030
Right-sided Pr >= F	1.0000
Table Probability (P)	0.0030
Two-sided Pr <= P	0.0034

Table 9: Fisher's Exact Test Output from PROC FREQ

EVERYTHING ELSE

Being a successful statistical consultant means being a generalist. That starts with reading in your data. Almost anything can be read into SAS with a little ingenuity. This blog post (De Mars, 2020c) explains how to move data from an SQL database to SAS Studio with a dozen clicks. Love Excel or hate it, it's ubiquitous in the business world, so PROC IMPORT is going to be your friend. Need to read data delimited with a pipe, on multiple lines or with multiple records on a single line? Check this paper by Cohen (2010).

When I was a graduate student, Very Important Professors had lowly peon graduate students and programmers to write their code for them. All of those people had started their careers using punched cards, (honest!) it was that long ago. All of the statistical consultants I know write code, or, at least, can code their own analyses if necessary. Even if you aren't doing it all yourself – I'm certainly not these days – you need to know enough to review the code your minions wrote or help said minions when they get stuck. Sometimes, it's just quicker to do it by yourself than explain to someone else, especially if you need to fix a bug in a code that a client is waiting on.

In brief, even consultants who are well-known specialists in one language know and use others. If you think your career is going to be spent sitting on a mountain or in penthouse office, pontificating to others about sums of squares, the computation of Wilks' lambda or options for PROC GLMSELECT, you are going to be sadly disappointed.

CONCLUSION

This is a hill I will die on – it's much more important to your business and to science that data are coded and analyzed correctly than whether you used stochastic versus mean imputation for the 3% of records that were missing data. Yet, as a graduate student, I was required to spend a substantial amount of time computing regression coefficients with a pencil and paper and no time on how to best validate my data. Neither was presenting to a non-technical audience ever discussed. While an experienced statistician might doubt he or she would make such obvious errors as starting the INPUT statement in the wrong column or misinterpreting a logistic regression that used the DESCENDING option, I caught two of the mistakes in this paper this week. One was made by the smartest woman I've met in the last year. The other was made by me!

Aarts et al. (2015) selected 100 articles with high-powered designs from some of the most reputable journals in psychology. They were able to replicate less than half of the original results. This 'reproducibility' crisis is not limited to social science, with similar concerns being identified in chemistry, physics, engineering, biology and medicine (Baker, 2016). These non-replicated studies are in the highest tier of research, published in academic journals. How many more of the studies that failed to reach significance, had results that were difficult to interpret or ended up buried for other reasons had errors in coding or analysis? As a statistical consultant, you will often be the only person in the room with the expertise to spot those errors. Don't worry about making errors – because you will. Just try to catch them before you hit that 'send' button on the email with the attachment.

REFERENCES

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, urn:issn:0036–8075.

Aboumatar H, Wise RA. Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial. *JAMA*. 2018;320(22):2335-2343. *JAMA.* 2019;322(14):1417–1418. doi:10.1001/jama.2019.11954

Ange, B. , George, V. , LaLonde, D. & Wasserstein, R. (2019). 2018–2019 Academic Salary Survey. https://magazine.amstat.org/blog/2019/06/01/2018-2019-academic-salary-survey/

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nat. News* 533, 452

Cohen, M. (2010). Reading MORE difficult data. https://www.lexjansen.com/nesug/nesug10/ff/ff02.pdf

Crews, F. (2017). Freud: The making of an illusion. Metropolitan Books: New York City, NY.

De Mars, A. (2020a). AnnMaria's blog: Words from the prez. https://www.thejuliagroup.com/blog/

De Mars, A. (2020b). Data quality macro explained. https://www.thejuliagroup.com/blog/data-quality-macro-explained/

De Mars, A. (2020c). From PHPMyAdmin to SAS Studio for lazy people. https://www.thejuliagroup.com/blog/from-phpmyadmin-to-sas-studio-for-lazy-people/

Haan, M., Rice, D. P., Quesenberry, C. P., and Selby, J. V. (2011) Kaiser Permanente Study of the Oldest Old, 1971-1979 and 1980-1988: [California]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/ICPSR04219.v2

Hall, P. H. & George, V. (2017). 2015 Salary Survey of Business, Industry, and Government Statisticians. https://www.amstat.org/asa/files/pdfs/YCR-SPAIGsalarysurvey15.pdf

Kennett, R. & Thyregod, P. (2006). Aspects of statistical consulting not taught by academia. Statistica Neerlandica (2006) Vol. 60, nr. 3, pp. 396–411

Muenchen, R. (2020). http://r4stats.com/

SAS Institute, Inc. (2013). The CALIS procedure. In SAS/STAT® 13.1 User's Guide. Cary, NC: SAS Institute Inc. pp 1154-1858.

Schreiber-Gregory, D. N. (2017). Multicollinearity: What Is It, Why Should We Care, and How Can It Be Controlled? Proceedings of the 2017 SAS Global Forum. https://support.sas.com/resources/papers/proceedings17/1404-2017.pdf

Sen, S. (2020). Estimating percent effort. AmStat News, 512, 12-13.

Sijtsma K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

The Analysis Factor (2008-2020). The Analysis Factor blog. https://www.theanalysisfactor.com/

ACKNOWLEDGMENTS

Support is gratefully acknowledged to the U.S. Department of Agriculture Small Business Innovation Research program of the National Institute of Food and Agriculture

CONTACT INFORMATION

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

AnnMaria De Mars
The Julia Group
annmaria@thejuliagroup.com
www.thejuliagroup.com Twitter: @annmariastat