The Data Analysis Workflow

Vernon Gayle

For NCRM Online Resources https://www.youtube.com/watch?v=ZihBzaU2fF0

The data analysis workflow. I'm professor Vernon Gale, I'm professor of sociology and social statistics at the University of Edinburgh and part of the National Centre for Research Methods. At the current time the National Center for Research methods are unable to provide any face-to-face teaching. I hope that you and your families are all healthy during this difficult period. This presentation and its associated resources are all about the data analysis workflow. A thought experiment. Be honest, have you ever lost a file? Have you ever wondered if you've deleted a file? Have you and a colleague ever been working on different versions of a file? Have you ever struggled to identify data files? For example a file called chapter1_2019.dat and a second file called chap1_2019.dat. Well if you experienced any of the above you could improve your social research workflow. The examples in this presentation tend to lean towards statistical analyses of social science data but many of the issues around the work flow and being well organized at permeate across a number of research processes so it's worth continuing to listen in. Here's me one Saturday morning I looked fairly frazzled I'm working fairly hard at this stage but if I'd improve my workflow it could be a different picture. Here's me on alternative Saturday morning this is the west coast of Scotland believe it or not at a time before Corona when we could still go outside. Workflow, what is it. Well for me it's the whole process from conceiving of an idea right through to its publication.

Commonly you'll download some data, here is the National Archive at Essex University, you'll download the data you'd be using your computer or your over your laptop or a machine on your university network you're probably using

a statistical software or programming language so something like SPSS, Stata R or Python and the first stage will be data wrangling. Next stage would probably be something like exploratory data analysis followed by something more formal and statistical such as statistical modeling then these results will be written up and submitted for example to an academic journal. Then after peer review hopefully they'll be published in the journal. For many of you the process will look the same but the output might be something like an Msc dissertation or your PhD thesis. Analyzing data without a planned and organized workflow can be compared to drinking and driving in both situations it doesn't matter how careful you are it's still highly likely to end in a wreck. These are the wise words of Professor Philip Stark UC Berkley. Therefore, just like drinking and driving we strongly warn against not having a systematic workflow. The workflow. The work flow should be planned and carefully orchestrate. The workflow must not be ad hoc so it mustn't be worked on piecemeal and it mustn't be developed in reaction to mistakes. The work flow cycle is relatively straightforward to conceptualize. Plan, organize, compute and then document. After gaining access to your data, having downloaded it, this is a picture of the UK data archive at Essex University, you probably have your data on a laptop or on a machine on your university network and you'll tend to be using a statistical software such as SPSS or Stata or programming language such as R or Python and you'll begin your data wrangling.

The first bit of advice is don't use drop-down menus, I'll say it twice, don't use drop-down menus. If you use drop-down menus you'll have no audit trail. The audit trail is nothing more than a line of breadcrumbs that lead you back to where you started and you won't have one if you use drop-down menus. Write out your data wrangling commands in syntax using GUIs Graphical User Interfaces will leave you in a sticky mess. What does the data wrangling phase involve. It involves a number of steps commonly selecting variables surveys tend to have a large

number of variables but we will only want a small number of them for our specific analyses. We need to operationalize measures the data may have information on income it may be net income or gross income or household income and we need to operationalize the measure given our research question. We'll need to recode variables to get them into the specific format we need for the data analysis that we're going to undertake. We'll need to often select cases so surveys tend to be large in general we may be only interested in the subset of cases for example just married couples or just households with people from certain ethnic minority backgrounds and so on. And inevitably any data set will have missing data of missing information and we'll need to know think of hard and work out how we're going to deal with that information. The question is can these actions be traced in my audit trail. Every action that transforms the raw data into the analytical data set should be able to be traced within the audit trail. Minor actions in the data wrangling phase can have major consequences later on.

After we've transformed our raw data into a analytical data set in the data wrangling phase we tend to go forward and do some exploratory data analysis after which we go forward and do some more formal analysis usually something like statistical modeling. But it's not a straight linear phase you quite often have to loop back around and do some more exploratory data analysis we sometimes have to go further back and do some more data wrangling as well and then we end up with a set of results that we'll write up and ultimately submit for peer review. In the data analysis phase there are many many operations that we routine undertake that are very easy to overlook.
For example which cases did we choose in an analysis or the format of the variables, how do we treat missing data, which estimation method that we use you may be estimating the model and we may have to decide between using maximum likelihood estimation or generalized least squares for example. We may have to think about

which weights suit our analysis or how we're going to represent the structure of the survey. We may be doing something exotic like bootstrapping and we need to think about setting a seed for generating random numbers. We may be doing something like fitted a random effects model and we are trying to decide the number of quadrature points that we use the standard or the default number or do we need a different number of mass quadrature points. We may when working with software such as R or Python have to think about which library to choose to undertake a certain procedure or indeed which version of a software package we're using and often many of these operations are overlooked but they're vital parts of the data analysis phase. The question again is can these actions be traced in my audit trail and once again every single one of these actions should be traceable in the audit trail.

Improving the workflow. The workflow should better support you and what you do. Not changing you into something you're not. So really what you want to do is improve your workflow to support you and how you work rather than implementing a kind of very regimental very kind of Stalinist almost view of how you should work. Because if you do that then you won't be able to stick to it. So you need a workflow that better supports you and what you do. Remember all the serious work must be reproducible there must be an audit trail and I'll return to this point again but I'll say it again now. All serious work must be reproducible there must be an audit trail. A planned workflow has a number of benefits. I sometimes call these the four pillars of wisdom. They are accuracy, programming efficiency, transparency and reproducibility. What do these mean.

The four pillars of wisdom. The first is accuracy minimizing information loss and errors in analyses and outputs. Programming efficiency, automation for example maximizing the use of features in software. So for example rather than

doing something 10 times for 10 waves of data you need to do something like write a loop that will loop over the 10 waves of data so you're using the software the programming language to work more efficiently. Transparency, you need to be able to show what you did, why you did it, when you did it and how you did it. It's a transparency showing what you did why you did it,
when you did it how you did it. So this leads on to reproducibility can you get the same result every time whoever is he's doing it or wherever it
is you're doing it do you get the same result if you run the analysis on your laptop or on a university network machine. The work should be reproducible.
And this will help, especially when editing, this is a point I'll come back to so when you're rewriting things like rewriting parts your thesis having to undertake new analyses for example or when you submitted the paper and you've got evil referees comments back and you have to go back to the work and do more or change things. So this this will be a natural benefit.

I tell you Long's lawn out his J. Scott Long from Indiana he says it's always easier to document today then it is tomorrow.
Corollary number 1 of Long's law nobody likes to write documentation.
Corollary number 2 nobody ever regrets having written documentation. And the final thing long says is has anyone in the history of data analysis ever said these files are too well documented. Keep calm and write comments. So I'm going to get you to visualize for a moment it's about 4:15 on a Friday you're working on something you need to go relatively soon in fact members of your department are meeting up for a drink after work what do you do. Right it's very very tempting to just save your file and think I'll put some comments in first thing on Monday morning well that's a good intention but we can all guess what's going to happen when you come in on Monday and there's another 40 email to answer and students waiting to see you for example yeah if this is all going to fall apart. Write comments however cursory comment, comment, comment.

This is the key to a good workflow, is having lots and lots of commentary lots of narrative in your documentation. But the good news this is a good news story improving the workflow can be done with a modest amount of effort. And the other bit of good news is the less experience you have the better because you can just start from scratch. If you're a very experienced researcher one of the things you can do is say wake up and say that from tomorrow every new project you're going to engage in you're going to improve your workflow but if you're starting out if you're a master student or a PhD student or a early career researcher you can just say like from now on from tomorrow I'll wake up and I'm going to try and follow good workflow practices. At this point you may be sitting there thinking please give me some practical strategies and tactics and tips for improving my workflow and there are a number of things you can do quite easily that will help your workflow. Having a standard directory structure will help you enormously in your workflow. Here's one that suggested by J Scott Long it's quite elaborate, it's got a code books in one folder clean data in a folder, raw data another folder, Stata do files in another folder and other folders for documents, figures, logs, tables a temporary folder that could be used for example when merging files or matching data, a trash folder clearly for things you don't want and a working folder that you may be using for kind of day-to-day working before something is lodged more permanently in one of the other folders.

File naming conventions are very important that it takes us back to the start of this presentation when I said have you ever wondered about a file. Here's another J Scott Long suggestion one that I've worked with for many years now so a file name for me has a name, it's normally something that is very eye readable so something that will give me a clue as to what it's all about, a date the depositors initials, the version and the type. And the versioning is particularly important in terms of

keeping track. So for example British household panel study BHPS the wave A individual response file you can see that from the eye readable name and that's their protocol down at ISER, It's a very good one so the BHPS wave A individual response so that's bhpsaindresp then I've got the date of the file so I've got the year the month and then the day. You'll produce many files in a year so having it round this way is very important because you don't want to have loads of files ending in 2014 I have produced some other file since 2014. So I've got the eye readable name, the date the depositors initial that's me VG and the version so it's v1 and this is a .dta file so it's a Stata data file. So this would be a bhps file aindresp it's deposited on the 6th of May 2014 it's deposited by Vernon Gail VG and it's a version 1. And the file type is a Stata.dta file. So a good file naming convention again you can start from tomorrow if you do so it will help you enormously certainly there'll be a time in the third year the final part of your PhD might be late in the third year even where you're looking for files produced in year one and this sort of protocol will save you lots of time lots of anxiety lots of stress it will help you keep track.

The analysis file depending on what software you use if you use SPSS it will be a syntax file if you use Stata it will be a .do file if you use R to be an R script it might be you may be working in R markdown you may be using R studio if you're in Python for example when I work in Python I tend to use Jupiter notebook but whatever system you use whatever software or or data analysis language you're using, whatever you're using have a system so that when you write out your syntactical commands both in the data wrangling phase and the data analysis phase you have some general information. So for me this is the sort of information that might come at the start of a file so information on the author, the project, the sub-project it's part of, the date of the next meeting or my next supervision, what the date of the latest

updates is and a track of the previous updates. And you notice I have my next actions at the top of the file. What lots of people do is they write some kind of note to themselves about what they've got to do next but that comes after a thousand lines of code so they have to scroll all the way down.
When I open up one of my files I can see what I'm supposed to do next right at the top. So that kind of headlining helps me understand where I've got to on what I've got to do next. Variable naming protocols. Good data providers will tend to stick to clearly defined verbal naming conventions so if you've downloaded a data set like the PSID from the US or the SERP from Germany or the British Household Panel Study or the UK household longitudinal study or any of the British Birth Cohorts the data providers will have done lots of work to curate the data and they'll tend to have fairly good naming conventions. This for example is wave one for Understanding Society the UK household longitudinal study.

There's a variable, gender variable called a_sex. So in wave A of the survey when one of the survey is _sex in wave 2 its b_sex and so on and if you go to the UK Data Archive and you'll see this in study number 6614. Similarly if you go to the Youth Cohort Study Time Series for England Wales and Scotland this spans 1984 to 2002 the variable t0schtyp school type once again eye readable very very useful and so t0 time 0 5.0 it's a type of school that the people attended in year 11 and so it's at time 0 of the survey and if you go to the UK Data Archive this is in study number 5965.However look at this survey question. It's a survey question it says here are some things both good or bad which make people which people have said about their fourth year and fifty fourth and fifty years at school we would like to know what you think please tick a box for each one to say whether you agree or disagree. School has helped to give me confidence to make decisions. And the person either the pupil

either agrees or disagrees. In the data set this is what it looks like it's by contrast the youth cohort study of England and Wales and cohort 4 this is which is study number 3107 this variable hair is given the opaque name dx11_a. Not particularly eye readable so you can't guess what it could be and you know how you'd never get some information about that. So again a protocol you know think about your protocols this isn't a particularly use what the useful one for naming variables in the dataset. I'm gonna move aside for a moment and just talk about estimating work time just so that I say this and it because it's often overlooked. I went to Sterling University for a long time I worked my colleague and personal friend Professor Paul Lambert and the students there quite often joke that it was something called the Gayle Lambert constant the idea that work takes five times longer than you think it does. And I actually got some people to test this a load of data analysts in the last couple of years to estimate the time that certain tasks would take and really the five times constant is very very very very close to being correct a lot of the time.

So essentially if you think a piece of work is gonna take an hour it'll probably take you five hours if you think it's gonna take a day it will probably take you about five days if you think it's gonna take you a week and it's probably take you five weeks and so on. Why is this important well this is very important when we kind of plan work but it's also importantly kind of when we you know decide to turn work into supervisors when we estimate how long things are going to take it becomes absolutely critically important for example when writing grant proposals and costing data analysis when doing consultancy work and so on so I would say data wrangling data analysis always takes five times longer than you estimate so think really hard about that and it's five times when you've got a good work flow. So you can make it much longer than five times by having a kind of shoddy work flow but I think you can

probably really get in the habit of thinking
about how long the tasks gonna take and then thinking well I need a slot that's
about five times that to do a good job even with a good workflow. The benefits
of a good workflow. There were a number of benefits to having a good workflow.
First of all it limits the duplication of effort a good example is if you've
got a good workflow and you've analyzed a single wave of a hassle panel study
for example this will allow you to analyze subsequent waves without
duplicating as much effort. The next is systematic work minimizes errors and I
would say that systematic work in most walks of life will minimize errors. It
also helps us to detect errors.

It becomes critical when editing work whether this be revised in thesis chapters or revisions that are made after evil referees have sent back comments and sometimes these require
quite a lot of work one of the things we know about the publication process is
sometimes there's a large lag between undertaking the work and submitting it
and then get in comments back and if you can rapidly return pick up where you
left off and also go back in the work phase the the workflow process and
sort things out this will be a major benefit. It will aid collaboration so if
you're working across teams or into are doing interdisciplinary research or
working with people in another university or even in another country
having a clear workflow so that they can look in see what you've done add to it
change parts of what you've done and you can understand what they've done and
change parts of work or add to things that will aid your collaboration and
speed up outputs from it. It will support additional work in the future. So whether
that be work that immediately follows on or new work that you take up that draws
on the work that's previously done having a good workflow that is well
documented will always help to support additional work.
This researcher here was aided by an early career researcher it was this
researcher here. Help your future self by having a good workflow. What to do. First

always have an audit trail. Second don't use drop-down menus a GUI will land you in a sticky mess. Write out the commands you need for data wrangling operations and the commands you need for data analysis operations. Have a systematic directory structure.Have a convention for filenames and for variable names and pay attention to version control, that helps you keep track of things. Finally most importantly - plan. Don't do ad hoc work or work on the fly and write comments you can never have too many comments write comments, comments are your friends. J Scott Long produced a book called the workflow of data analysis using Stata it's a real Bible on the subject of the workflow and whatever age or career stage you're at it's worth a read he has also posted a really good PDF version of a talk on the work flow the link is below and a few years ago myself and my colleague Paul Lambert wrote The Workflow a Practical Guide to Producing Accurate, Efficient, Transparent, and Reproducible Social Survey Data Analysis and that's an NCRM working paper that's available from the address here. Remember all serious work must be reproducible