Networking

We have a mailing list! This is an effort to create a CAMH community that is interested in all things Statistics, Data, Evidence and Research in general. To subscribe to our mailing list, click here or contact Marcos. And feel free to mention the mainling list to your coleagues.

Currently our mailing list is used to broadcast our weekly update (Stats News), where we include information that may be of interest to researchers at CAMH, and that is related to data or statistics. That includes education posts, papers and resources, Software tips and scripts.

We have a Statistical Education section where we talk briefly about a new topic each week, usually related to the type of biostatistics that researchers at CAMH use.

Stats News Education Posts

062 (23/June/2020) - Independent Sample t-test *to Top*

We have received a couple of questions about the use of independent sample t-test lately, and so I thought I should try to address it here shortly.

The Independent sample t-testis is a test that we use to compare the means of two independent samples, and by independent we usually mean that the two samples are composed by different subjects in our setting at CAMH. For example, we want to test if smokers and non-smokers have the same average depression scores. This is the typical situation where one would use the independent sample t-test.

However, I think it is fair to say that nowadays the t-test is unanimously recommended against in pretty much any practical situations. That is because the t-test is optimal and therefore better than other tests only when its assumptions are met. In particular one of the most damaging assumptions is that of equal variance in both groups (using our example, that the variance of the depression score is the same among smokers and non-smokers).

The alternative to the t-test in the situation where variances are not equal (homogeneous) is to use the Welch's t-test. R will do that by default if you use the "t.test" function. SPSS will also show it in the "equal variance not assumed" line of the t-test output. However, SPSS will also give you the result of the Levene' s test, which tests exactly if the variances are equal. We are then lead to use the "equal variance not assumed" p-value only when the Levene's test is significant, which is not quite the best thing to do. This is because a non-significnat Levene's test does not mean evidence of equal variance, it means lack of evidence that the variances are different. And in particular when you are dealing with a small sample size, which is many times the case, such non-significant output means very little because you lack power to detect differences in the variances.

So, basically what we recommend is that you just always use the Welch's t-test, the one that does not assume equal variance.

There are other assumptions made by the t-test and Welch t-test, like normality and not having outliers, and you being interested in the mean. If your data is visibly non-normal, you are probably better using the Mann-Whitney U test, the non-parametric version of the t-test. I also like the Kolmogorov-Smirnov test in that it compares the distributions rather than simply the mean or median. When you have outliers, tests like MW U test is really nice as it is based on ranks, not on the actual values of, say, depression scores.

If you want to read more on this and some references, along with some SPSS and R output, I would point you to a Daniel Lakens post relative to this subject matter.

061 (16/June/2020) - Randomization Tests *to Top*

Index of Stats Education Posts

The latest Tutorial in Biostatistics published in Statistics in Medicine, shows how to conduct randomization tests for randomized controlled trials. We will shortly go over randomization tests here and you can consult the paper for more details and references.

Randomization tests are an alternative to the parametric tests based on distributional assumptions that we use when analyzing our trials. And these parametric tests are pretty much all we do and it is difficult to see a clinical trial analyzed with a randomization test. By parametric we mean that we rely on distributional assumptions, like normality, so that our test statistics (like the t statistic) and its p-values are valid. If the data is normal then the test statistics follow a t-distribution under the assumption that "the null hypothesis is true". And that is nice because if we calculate the said t-statistics and it does not look like it came from a t-distribution then we infer that our assumption that "the null hypothesis is true" was violated. And we say that we have evidence that the null hypothesis is not true, which usually means that we have some effect.

With randomization tests this reliance on distributional assumptions of the data and of the test statistic is not necessary. Instead your assumption is that if there is no treatment effect then you would get similar effect size under a different randomization (that is, if you re-assign groups to subjects). If there was no treatment effect, then the treatment assignment does not matter. So, we randomly re-assign treatment to participants and re-calculated our statistics which could still be the t-statistics, or it could be just the average treatment difference. Ideally you would re-assign treatment groups to subjects in all possible ways, and calculate the mean difference each time. But even with not very large sample size that would require a long time since there are a huge number of ways to re-assign participants to treatment groups. So, what is done is a Monte Carlo simulation where you re-assign treatment groups to subjects a million times, say, and that should give you a good approximation to what you would get if you had covered all possible re-assignments.

If for each re-assignment you re-calculate the mean difference, then you can compare the distribution of the differences with the difference that you got with the original data. If there is a clear effect in that a group has better outcome than the other then what happens when you re-assign treatment groups to subjects? Well, you basically destroy that difference most of the time and you will find very few if any instance where the difference is larger than the one you got with the original data. The proportion of times that happens is the randomization p-value.

The paper also describes methods for creating confidence intervals and to analyze binary data, but for now we will stick to conveying the main idea. In their Table 4, on the right side, they provide a quick algorithm for the randomization test which we will reproduce here. The Delta in the steps below is related to the difference we expect under the null hypothesis, which is usually zero. The N is a large number, I would say 10,000 or more.

(1) Generate a new treatment assignment sequence.
(2) If a patient in group A was re-assigned to B, add to the outcome.
(3) If a patient in group B was re-assigned to A, add to the outcome.
(4) Calculate a new estimate for .
(5) Repeat (1)-(4) for N times, estimate a P-value.

Finally, I just want to notice that the randomization test should be well accepted by journals, I don't see why they would not like it. You need to be a little careful in proposing randomization tests if you plan to run many tests, which will make things time consuming for you, and things like Bonferroni adjustment would be equally applicable for randomization tests because the goal is still to calculate p-values .

060 (09/June/2020)– Categorizing Continuous variables *to Top*

A common practice is to categorize continuous variables and use these categorized versions in statistical analysis. For example, instead of using BMI as a regression predictor, we may split it into categories, like Obese/Not Obese. Such approach will pretty much always have a cost in terms of power or precision, meaning that your statistical models are not going to be as good as they could be. In what follows we list some reasons for not categorizing continuous variables.

1. Maybe the first point is just that we want to represent well reality, and we can deal with continuous variables in statistics, we don't need to categorize them if they are not originally categorical. Unless the reality is in fact binary we will be doing a poorer job modeling the real phenomena if we use a binary version of a continuous variable.

2. The categorical version of the predictor will be a less then optimal predictor. Lets say that you want to use BMI in a model to predict risk of X (X can be anything you like) and you decide to categorize it in BMI > Y and BMI < Y (say, Obese/Not Obese). There are two points here that we can mention. The first is that almost certainly that there would be a cut-off value Z not equal Y which would provide better predictive performance, and that happens because Y was chosen based on some clinical definition of Obesity, not on the optimal cut-off to predict X. The other point is that the cut-off value Y will almost certainly vary depending on other predictors in the model. For example, it could depend on age.

3. There will be a loss of power, that is, you will find it more difficult to find effects. This can be easily shown in simulations.

4. It implies an assumption that is not needed: that the relationship between two variables is flat in the intervals that were grouped together. For example, if instead of continuous BMI we use Obese/non Obese and measure the association with the condition X, we are assuming that this association is the same for all BMI values in the Obese range (BMI > Y) and in the Not Obese range (BMI < Y). This is unlikely to be true and in fact we may miss some interesting association between BMI and condition X that may happen away from the cut-off point Y.

5. Related to the above, the reality is that the actual association is always smooth, it does not have a jump at some points. For example, if you categorize BMI into Obese and non-Obese, then you look at association of such binary variable with condition X, you assume that exactly at Y there is a jump in the association. That is never true. If BMI is associated with risk of condition X then there will be a smooth association and the risk will increase as BMI increases, or behave in some other continuous pattern, it will not only increase at Y, so that Obese folks has a risk, non-Obese has another risk.

6. Usually, there is no good definition or reason for choosing a given cut-off. Different researchers may disagree on what it should be.

7. Categorization may invalidate p-values related to association tests, and in many instances, it reduces credibility. If you decide on the value of your cut-off for categorizing after looking at the data, you must make sure you don't define the cut-off based on what you see because you may be choosing a cut-off that appears to be associated with more group differences. Lots of this goes on where folks try to choose cut-offs that can provide a significant association. In that sense, if you don't have very good justification for your cut-offs so that they were defined a priori, the mere use of cut-offs may cause your research to be questioned and lose credibility.

If you are categorizing a continuous metric because that is what is usually done (sometimes we do that just so that results are comparable to what has been published), consider also doing the analysis with the original version of the variable and making those results available as well.

If you are trying to address non-linearities, I would suggest that you consider non-linear models like splines or simple smoothers. You can also try simple non-linear functions, like a quadratic curve. Time is one of the few variables for which we usually have good justification to use it as categorical instead of continuous, usually because time was measured in a categorical way even though it is a continuous variable: it is not like you can access your patients at every minute, for example.

I will finish by leaving you with a couple of references, in case you need them to make a case for using continuous version of some variables:

Streiner, David L. "Breaking up is hard to do: the heartbreak of dichotomizing continuous data." The Canadian Journal of Psychiatry 47.3 (2002): 262-266.
Senn, Stephen. "Disappointing dichotomies." Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 2.4 (2003): 239-240.

059 (02/June/2020) - The SIR compartmental model for spread of disease *to Top*

Our text today is based on a text published in the JAMA Guide for Statistical Methods this week. For a more detailed explanation of the topic we recommend that you take a look at the paper. This topic is deep into the field of epidemiology and to the extent that mental health conditions are usually not contagious, the SIR model is not the first thing that comes to our mind. However, it offers a nice example of a model and its assumptions, which may give you the concepts you need to then use other more relevant models.

The above said, there could be some specific areas in which we could explore SIR models for mental health. For example, this text published in Nature looks at the spread of suicide ideas caused by the awareness of those ideas as they are spread on human networks. So, a person may have suicide ideas and talk about those ideas which may increase the propensity of others having these same ideas. Maybe we could also think of the SIR model being applied to the spread of substance use in human networks.

SIR stands for (Susceptible Infected Recovered), as the three compartments where individuals of the population may be at any given time. For example, in relation to Covid, most of us seem to be in the Susceptible compartment. The model has two parameters: the rate to which people move from Susceptible to Infected, and the rate they move from Infected to Recovered. The ratio between these two rates is the famous R0, which is the average number of individuals infected by an infected individual.

Although you can work with these models mathematically, nowadays it tends to be very easy to just run Monte Carlo simulations. As a result, you will get the number of individuals in each compartment at each time point. You may not be happy with that if you are into stats, you may also want some measure of variance of the results. Simulations will also allow you to get that as you can account for the variance and distribution of the parameters. In the case of Covid, for example, parameters can be very imprecise. A SIR model will usually be presented as a graph with three lines, each representing the number of individuals in a compartment at a given time point.

The SIR is a very simple model, which means that it relies on many assumptions and is particularly susceptible to criticism. And assumptions is a very important component of models in general. The SIR relies on several structural assumptions (regulating the extent to which the model actually describes the spread of the disease), while statistical models will also rely on inferential assumptions (regulating the extent to which the inferences are valid). These terms, structural and inferential, were created by myself just to try to make things easier.

The point is that the inferential assumptions in statistical models (normality, homoscedasticity, independence, influential points and outliers, overdispersion…) seems to often make people forget about the structural assumptions (linearity, causal links, functional form of the model) so that we often think we are good if our data is normally distributed and there are not outliers, or things like that. But the structural assumptions are more important, in my view. If you model something as being linearly associated with an outcome, when the correct functional form is not linear, you may got “valid” statistical inferences for a model that does not represent the reality.

In the SIR model, the assumptions are crucial, and there are several assumptions that are important. For example, you assume that once you are infected you immediately infect others. That is not true for Covid, for example, which has a latency period. You assume that the population is closed, but that is clearly not a valid assumption in today’s world where people travel everywhere. This is an assumption that while clearly not true may be good enough. You assume that a recovered person does not get infected again. You assume that there are no interventions (like social distancing and masks). You assume that the rates are homogeneous over the entire population (but maybe in some communities people are closer together than in other so that the spread could be fasters there).

I just want to say that there will always be assumptions, our models will never be perfect. So, the assumption is not a weakness. The problem is whether these assumptions are valid, and if not, what the consequences. Yes, because there can be good models with assumptions that are clearly bad, but it is important that you know, or have an idea, of the consequences of bad assumptions. Assumptions are related to math, so this also says that it is important that you care about the math behind the models you use.

In the case of Covid I tend to be suspicious about a model like SIR because it is not only too simple, but also requires many assumptions that seems to be quite important. In the case of the rate of transmission, we are quite in the dark and are not able to get it very precise since we don’t get to know about a large number of individuals who got infected but did not come to be tested. However, even in the case of Covid, the SIR model may be used as a nice and useful starting point in the pursuit of a reasonable model. The thing about Covid is also that parameters change all the time with the new interventions by the government, and geography, so that we would have to keep adjusting the model for the new and diverse scenarios.

So, we will leave you with the final point is that in the SIR model as well as in statistical models, you have to make the assumptions clear, and specify why you are making them. Try to test or argue about how consequential they are. That will allow for others to understand that you are making assumptions and also it will allow readers to access their perceived level of evidence in your findings, fairly considering the assumptions used.

058 (26/May/2020) - The SAP (Statistical Analysis Plan) *to Top*

This is a quick text to remind you about the Statistical Analysis Plan (SAP). You can find a nice article about the SAP in JAMA. So, the purpose of this text is to make you aware of the SAP so that you think about it before your next clinical trial, and that may make your life easier when you try to publish your results since top journals are requiring it more.
There has been increased pressure for us to have a study protocol, preferably one that is created before our data is even collected. This is more crucial for clinical trials, but I would argue that it is also important for observational studies that involve testing hypotheses, including the ones for which data is sitting there, already collected. At that point, before looking at the data, we should be able to define our main hypothesis and how we expect to test it in terms of statistical analysis. Once these things are in the study protocol, we then need to follow it, and document any departure from the pre-specified protocol, if at all. This will increase the credibility of statistical tests that uses p-values to decide if an effect is significant or not. Without a study protocol, a p-value is almost useless as far as I am concerned, and at the minimum it loses its interpretation as a probability.

However, nowadays some journals are requesting more than the protocol, they are requesting the SAP, which is supposed to be a document separate from the study protocol. This happens, for example, if you want to publish in JAMA or NEJM, and I am sure others. And it is mostly for randomized clinical trials.

I would say, though, that any data analysis benefits from a study protocol where the statistical analysis is well defined before you look at the data, and it will be stronger if it has such protocol. Not only that; to write down the description of the statistical analysis will help (force) you to define well what you want to do, and that will help you if nothing else. It is not uncommon to see folks doing their stats without much of an idea of what they want to do.

Back to the protocol and SAP, the idea is that the study protocol will talk about the statistical analysis in broad terms, but the SAP will be very detailed. In the previously mentioned guideline for the SAP, published in JAMA, they have a list of items that they recommend to be in the SAP. We are not going to go over those items here, but you can take a look and see that it is very detailed. Maybe you will find it too detailed. I have looked for SAPs published in JAMA and found that they (the ones I found, like two or three) are not so detailed, they don't really follow this guideline. So, you don't need to worry that you will need to follow such detailed guideline.

The above said, the details tend to be important and I would defend that they should be provided if you can. After all, you are conducting a serious and important clinical trial. Unfortunately, it is not difficult to find protocols where the statistical analysis is not completely specified, leaving room for multiple analyses being conducted where all are consistent with the protocol. That is detrimental to the final quality of the evidence as well as for reproducibility.

In general, if you do specify your statistical analysis with all the details in the study protocol things should work fine, however, by considering the SAP as a separate and specific document, you will cover more basis and ensure you are following the best practices.

That is it for today, the take home message being that you think about the SAP when planning a trial, think if what you have in the protocol is detailed enough, because when trying to publish your results you may need it.

057 (19/May/2020) – Change in Scale and Variable Standardization *to Top*

Two weeks ago we talked about the role of variable centering in Regression Models. Today we continue that discussion by taking a look at the role of variable standardization and changes in scales in regression models. When we talked about centering and now we are focusing on regression predictors (independent variables) and not outcomes (dependent variables).

By changing the scale we mean to multiply or divide regression predictors by some value. This should not change the inferences about statistical significance of the independent variable.

A nice example of a change in scale is, for example, if you divide a variable Age in Years by 10. What happens is that now you have Age in 10 years, not in years. As a result, the expected effect of age will be 10 times larger in a linear model. This can be helpful because the effect of a single year change in age is often very small, and it is kind of awkward to report. Like, the coefficient of age was 0.0001 (SE = 0.000022). No, don't report that, instead re-scale age! You will be able to get rid of one zero if you report the coefficient of Age in 10 years. And often it just makes sense that we don't expect things to have relevant effect over 1 year and so 10 years become more reasonable anyway.

If instead of age we have income in dollars, then besides very small coefficients you may even run into convergence problem for some models and software. Say, what would be the effect of increasing 1 dollar in one's annual income in terms of the probability of developing dementia? Such small increase in income does not even make sense, but it is what you are estimating if you use income in dollars. The coefficient of income will necessarily be tiny and can even cause computational problems. In this case it is better to measure income in $10,000, for example, so that your coefficient will be the effect of increasing the income by $10,000. You do that by just dividing income by 10,000. Or you could divide it by 20,000 if you think that is more reasonable.

The above are just changes in the scale of the variable, it is like you are using a different unit. The two examples above is what I prefer to do when working with well-known variables, which are variables that people often understand easily.

Sometimes we work with more abstract variables, and we don't really have a rescaling that we would think of as reasonable. Maybe we are looking at some concentration of some marker, or amplitude of some wave, for example. It is not clear that dividing by 10, 20 or 500 is what makes more sense. In these cases, we may prefer to divide by the estimated standard deviation of the variable. If you divide income by 10,000 you interpret its coefficient as the effect of increasing $10,000 in income. If you divide by its standard deviation, you interpret its coefficient as the effect of increasing one standard deviation in income.

For ease of understanding, lets go back to age. Imagine that we are looking at low income households (between, say, $10,000 and $30,000 – I am making this example up, but bear with me) and we want to look at the effect of income on probability of being victimized. In this case it is reasonable to think that the standard deviation will be around $5,000, which is probably a reasonable change for this population where $10,000 may be unrealistic.

In my view, regardless of the variable we are talking about, a change of 1 standard deviation usually can be though as representing a meaningful, reasonable change, not too small or too large. But that is a rough interpretation, one you can thing in broad terms; you should not use it if you have a better definition of what would be a reasonable change, maybe a clinical definition. Also in broad terms you can think that for a variable in general, most of its values are in between its mean +/- 2 Standard Deviation, so they vary by 4 standard deviation and a change of 1 standard deviation is a sort of a change of around 20 to 25% in the possible range of the variable, which tends to a reasonable change.

Well, you need to be aware that statisticians will come up with these things if you don't tell them what is a reasonable change for which you want to estimate the effect size. It is like you estimate the effect of depression as measured by some depression score, on some outcome. If we just use the depression score we will find the effect of increasing one point, which may be a to small, meaningless increase. So, you can define a priori that we will look at increases of 6 points because that is what is considered a meaningful change in the literature. It will make the results more interpretable from the practical point of view.

The other reason to rescale variables using their standard deviation is so that coefficients become comparable. Say, you may have both age and income as predictors in a model, and you get a coefficient for age and another for income, but then you want to compare them to have an idea if the effect of age is bigger than the effect of income. You cannot just compare coefficients because the effect of age is relative to a change of 1 year and the effect of income is relative to a change of, say, $1000 in income. So, the coefficients represent the effect of totally different things and cannot be compared. However, if you divide age and income by their standard deviation before adding them to the model, then both coefficients of age and income will be measured in terms of the same change of 1 standard deviation in both variables. This is considered more comparable, if you really need to compare them numerically.

I am saying this because I am usually against such comparisons, I tend to think that as a researcher you need to understand the effect of age and income, then think about both and do the comparison relative to all you know. For example, even if the income effect is quite small it is usually more important because income is something we can modify, age is not. Modifiable or not is an external thing that you don't solve by dividing by the standard deviation, and if anything, such division will divert your attention from other things that may matter more, like whether they are modifiable.

If besides dividing by the standard deviation you also subtract the mean (center), we call it standardization, or z-score. In this case you also get the benefits of centering. Standardization will not make your variable more normal, which seems to be a common thought out there. Usually you don't need to worry about normality of regression predictors, only of outcomes, and it is the normality of the model residuals, not the outcome itself, that is assumed by the model..

I will stop there. I will repeat here that we are talking about predictors, not regression outcomes. Another thing you can do with predictors are transformations, like the logarithmic transformation. These can be done to get rid of outliers, to make the model multiplicative, or because it is somehow relevant for interpretation, and also to address non-linearities which is a very important thing and we may talk about it sometime...

056 (12/May/2020) – The meanings of Statistical Significance in clinical trials *to Top*

Today we will comment a paper that came out in JAMA this past week. We encourage you to take a look at the paper because they have some nice examples of real trials that help getting the point across.
I though this was a nice paper to comment on because it makes cleat that the interpretation of a significant or not significant result may vary widely. That is just another way to say that being significant or not means very little by itself.

The key point that is emphasized in the paper is the need to take into account the cost related considerations always involved when stating that something is significant or not (which is interpreted as we are making a decision about the effectiveness of the intervention). Not only monetary cost is implied here, but all sort of costs from which important ones are possibility of harm to patients and suffering, as well as use of resources in general (time, equipment, bed that could be used by other patients…).

I think we can summarise it by saying that statistical significance offer you a certain amount of evidence, but the actual amount of evidence you want for a given treatment depends on a diversity of costs. We should demand more evidence from treatments that we know cause harm, or tie down lots of resources, because we don't want to risk the incurrence of a high cost in exchange for no benefit.

In that sense a non-significant result may be enough evidence for us to change practice, if the treatment we are looking at is already known to cause some harm. We may then stop using such treatment. On the other hand, a non-significant result may still be inconclusive, and we may want to not disregard the treatment if it is something very cheap for which we see no unintended consequences, particularly if the trial design is not very good.

There are two important points that deserves further discussion in the paragraph above. The first is the considerations about using something like a treatment for which statistical evidences has not been shown. It sometimes has to be done because we don't have strong statistical evidence. These considerations are things like ethics and the right of the patient to being informed. We need to consider all of that, the patient must be informed about the evidence to the extent of possible and we need to be careful that the use of such treatment is not interpreted as the treatment being backed by statistical evidence.

The other point is that the meaning of a significant result also depends on the trial design. And here we can get ourselves into a long discussion which is not the goal today. So, we will stick to a summary, which is to say that the statistical status in terms of the significance of a result may mean very little if the data quality is not good, if the power is low, if the trail was not properly conducted, if the measures are not the best ones, if the pre-registration does not exist or was not followed, if the statistical analysis was improperly conducted, etc. Yes, there are lots of things you must look at before you interpret that p-value.

Besides the cost and the technical aspects of trials, another important aspect is the amount of external evidence, particularly if the threshold used for declaring significant is 0.05, which is a very low bar. You can interpret 0.05 as the probability that our effect is some sort of random noise. The thing is that 0.05 is still a high probability for things that are consequential, a risk we don't want to take. As an example, you can think that the probability of a fatality given Covid infection has been put at around 0.005 (0.5%). That is 10 times lower than our beloved 0.05 (5%), but for us, since the cost is too high (loss of life), 0.5% is not a low risk and so we will not take that chance by avoiding infection as much as we can.

So, yes, 0.05 is too high for consequential decisions and we will want more evidence which many times will come from external sources. This external evidence can be anything, ultimately, particularly for those into Bayesian Statistics, but I would say that the best external evidence is usually replication with independent teams and experiments. However, I can also see situations where the researcher is very confident about the treatment working, based on experience, related literature, animal experiments, theory and anecdotes. Sometimes you see a significant result and you say "well, no big deal, not surprising at all that this has a positive effect…". You say that because you have some sort of external evidence, and sometimes that is very important. Other times it will be clear that the significant result is weird, and you will be skeptical. That is lack of external evidence for you. Obviously, you can have some distorted, biased view of things, which has to be factored in too.

So, again, maybe you should take a look at the paper and the examples, but the take home message in my view is that you must always consider a lot more than whether something is significant or not, and making decisions solely based on the significance is going to be associated with poor science practice and poor understanding or statistics

055 (05/May/2020) – Models are part art, part science *to Top*

Last week we pointed you to some sources that looked at models for forecasting the spread of Covid, and then it occurred to me that we have an interesting example here of the fact that the modeling endeavour is not really objective. Otherwise, why would we have so many models? So, I just wanted to explore a little bit this point today as I think it is relevant for how we build models.

The first point is that there is no ready recipe out there for how to model your data. Although we will find the more traditional models to be frequently used in a sort of step-by-step way of model building, you not only don't need to cling to any specific model as if it is the right model for your data, but also it may be counterproductive to do so in the sense that it impairs your ability to do good, creative modeling.
In a more statistical-centric view we may say that there are bad models, and some models are clearly bad, but it is usually unclear what the best model is. And what happens is that experience and even creativity will play some role in the model selection and in our ability to find a good model.

As I said, there are not recipes as far as I know, and that is why I think we still don't have algorithms doing a good job in predicting the spread of Covid, or algorithms selecting models in general, for that matter. It is not that simple. In my view you will need a substantial amount of experience and knowledge of the different alternatives and of the subject matter in order to be able to build good models. The experience relates to the ability of visualizing the important aspects of the data, model and research question, all at once, and putting that together in a reasonable model.

The subjectivity in model building is also in the assumptions that we have to make, as is also well exemplified in the many Covid related models out there. For example, if a given model demands a specific epidemic spreading parameter, different researchers may reach different values for the parameters depending on their personal view of which data is best used for that end. Some folks may think that the available data is low quality to the extent that their personal guess may be more accurate, taking things to the extreme of subjectivity.

Another major source of assumptions and subjectivity in these models comes up when we must estimate or build models for the effect of physical distancing. Any mathematical definition of physical distancing will end up simplifying things to the extent that one must make decisions as to which simplifications are less harmful to the precision of the model.

Two weeks ago we talked about ARIMA models here in this section. These can be very objective models, and they are the kind of model that computers don't have much trouble fitting to the data themselves, without human help. However, I have seen zero ARIMA model for the Covid trend! It is important to say that all models we have seen are actually modeling a series of cases, so we could use ARIMA, why not?

Well, you probably could try to use ARIMA models to fit the seasonal flu series, because it is long enough by now, and more importantly, stable. But in the case of Covid, the lack of information is so much that objective and simple models like ARIMA would have no chance to do a good job. Experienced researchers don't even try it. There is also the fact that the ARIMA models are not traditional part of the epidemiological toolbox, so epidemiologists may not think of it because of that, but there are a lot more folks out there trying models than epidemiologists.

Those of us who are deep into math may not be very happy when we talk about adding subjectivities to our models. There are still folks against Bayesian statistics because it includes the subjectivity of the priors. But I guess we must recognize that the world out there is messy and our minds are able to add information to models that are not in the data. Information that comes in the choice of model and assumptions and our brains. And so the computer depend on us. For the time being.

054 (28/April/2020) – Centering predictors in Regression Models *t o Top*

This week we got a couple of questions about centering variables in regression models. This is a subject that is always around so we will comment on it briefly.

First, lets define centering. It is usually defined as the transformation that implies subtracting the mean from the original variable. Let's say that you have the variable Age. You first calculate the average age of your sample, say it is 35 years old. Then for each subject, you subtract 35 from their age. So, a subject who is 25 years old will become -10 and a subject who is 50 will become 15.

What happens is that now the average of the transformed Age will be 0. The Standard Deviation of the transformed Age will not change.

This should not change the model inferential results, in general, in that you will get the same p-values as compared with the non-centering version of the variables. However, if you have continuous variables involved in interactions, they may become easier to interpret, and in particular the main effect of variables involved in interactions will be interpretable when variables are centered.

Imagine that you have Age and BMI, and you centered age at its mean of 35 and BMI at its mean of 28. And you have a linear regression model with only these two variables predicting depression score (DS).

If the model does not involve interaction, then the model intercept is interpretable as the average DS for subjects with Age and BMI at 0 (that is, average Age of 35 and BMI of 28). If you had not centered Age and BMI, the intercept would not be meaningful because there is no subject with Age = 0 and BMI = 0 in the non-centered version of the variable.

If the model has an Age * BMI interaction, then the coefficient of the interaction is how much the Age effect changes when the BMI value changes from 0 (that is, 28) to 1 (that is, 29). The way I think of it (this may be helpful to some) is that if you make BMI = 0, then the Age main effect coefficient is just the Age coefficient because the Age * BMI interaction become Age * 0 and vanishes. If you make BMI = 1, then now you have Age main effect coefficient (you also have the BMI main effect coefficient, of course), and the Age * BMI is Age * 1, that is, the interaction coefficient became another Age coefficient. So, for BMI = 1 the Age coefficient is the sum of the Age main effect coefficient and the Age * BMI coefficient. You can then see that the Age * BMI coefficient is just how much the Age slope increases when you move from BMI = 0 to BMI = 1.

The above paragraph plays out basically the same way if we focus on BMI effect instead of Age, that is, the Age * BMI coefficient is how much the BMI slope changes when we move from Age = 0 (35) to Age = 1 (36). If Age and BMI were not centered, this interpretation would not be possible.

So, in short, if both Age and BMI are centered, the main effect of Age is the slope of Age for subjects at centered BMI = 0 (that is, its average BMI of 28) and the Age * BMI coefficient is how much that slope changes when you go to centered BMI = 1 (that is, BMI = 29). You can easily get a sense of how much the slope of Age (the effect of Age) changes as we change BMI (how BMI moderates the effect of Age), and what is the effect of Age for an average BMI individual (coefficient of main effect Age).

An important point to make is that centering does not have to be at the average, it can be at any value that is relevant. For example, I could center Age at 50 instead of at its average 35, if I am more interested on effects of variables at 50 years old rather than the average Age.

Most of the time the models I see at CAMH have no interaction and so there is not much point to centering predictors. That said, centering is hardly harmful, I mean, it may take you some extra time to do, plus you have to remember that you are working with centered BMI (so that BMI = 25 will probably not exist in the new scale), but those don't tend to be consequential things.

Finally, I want to discuss one thing which has been contentious, which is that centering helps with multicollinearity. That is the case only if you have interactions in the model, but you are still interested in the main effects. The coefficients of the interaction, and p-values, will not change whether you center the variables involved or not, but the coefficients of the main effects will. You will have in general more precise estimates of main effects in models with interactions, if you center your variables. However, most of the time you are really interested in the interaction, not main effects, and so, centering will not make any difference. Again I want to mention that if you center your variables, that will not harm you. A paper I like that talks about centering and multicollinearity is this one. As evidence of the contentiousness of the subject, someone criticized the paper and the authors published this short follow-up. These papers are nice if you want to get a little deeper into the subject.

We intend to soon follow up with a complement to this text where we want to talk a little bit about changing the scale of the variables (like when you divide by the standard deviation and standardize, rather than center, the variable).

053 (21/April/2020) – Time Series Analysis *to Top*

Today we will have a brief introduction to Time Series Analysis, which is not very common among the things we do at CAMH, and that being so we find that many folks are unaware of this statistical methodology, and sometimes mistake it for other techniques related to more usual types of longitudinal data analysis. Here we will explain a little bit of it, and in the next section we provide a nice resource. A nice R related text on Time Series Analysis in R can be found here.

Longitudinal Analysis is often said of the types of techniques that handle data correlated in time, but with few time points. Its goal is pretty much never projection into the future (forecast), as we have little statistical information for that with few data points in time. Instead, the goals of longitudinal analyses are most often to understand change in outcomes between specific time points. This can be done in different ways: by comparing outcome changes or trajectories between groups like intervention and control group; by just estimating changes in time, like from beginning to end of the trial; by gathering evidence for effect of moderators on outcome changes or trajectories; by studying patterns of trajectories; etc.

Although the term “trajectories” is often used in the context of longitudinal data analysis, these are not long sequences of data points, it can be just a two-time points trajectory, and we hardly see more than 4 or 5 time points trajectories. It is also often across many sampling units, which are usually human subjects (that is, you have many trajectories in a single data set, one for each subject).

Sometimes, though, you may face a sequence of data in time that is composed of many time points, say, 30 or more. That is when you think about Time Series Analysis. Because the cost of data collection repeatedly for the same subject used to be expensive, time series data is traditionally not collected at the subject level. For example, you can have daily number of Emergency Department visits as a time series. Or monthly sales of a certain medication. Or weekly number of traffic crashes in Toronto. Or the number of folks entering a subway station at each 10 minutes.

You can see that once you have a series of data points, the type of analysis you can do will need to be very different than the usual regression-based longitudinal data analysis. Even though regression models are so amazing that you can still use them for time-series analysis, specific models have been developed for this kind of data, namely, Time Series Analyses techniques. The usual main goal you have will also be different as it relates to future forecasts, not to specific data points and specific changes, except if you are interested in studying interventions at specific time points and how they affect the series. For example, one could be interested on studying the effect of cannabis legalization on a time series of monthly cannabis consumption.

Probably the most popular time series models are the ARIMA models. They tend to work well in many situations. The idea is to build in an equation a model for the current time t data point (dependent variable) that uses the previous time t-1, t-2, t-3… data points as predictors, as well as previous error terms e(t-1), e(t-2), e(t-3), … . Since data points that are lined up in a sequence tend to be highly correlated, so will be the error terms, and this correlation means that the error (residuals) may carry useful information about the next data point. You can see that in such models a point can be seen as being used twice, as the outcome and then as a predictor for the data points in the future.

ARIMA models can be extended to include seasonal effects as well as intervention effect. Imagine that you have daily visits to Emergence Department at CAMH as your time series data. A Time Series model will show you that the number of visits tomorrow is correlated to the number of visits today, possibly yesterday, possibly 7 days earlier (a weekly seasonal effect), and even maybe 365 days earlier (a yearly seasonal effect – this will be the case if you notice seasonal patterns in the data, like more visits in the winter or summer). In a given day you may add a reception desk to help visitors, and once you have a reasonable amount of daily data after such intervention you can test it, that is, see if adding the desk changed the series in any way. Besides the series of daily visit, maybe you also have daily average temperatures or rainfall, and so there are models that can be used to test those effects, including with lagged effect (for example, the temperature today may have an effect on the number of visits tomorrow).

Even though there are models that deals with many time series simultaneously, that has not been the usual case. However, with data collection becoming easier and more affordable, such situation is becoming more mainstream. As an example, you can think of wearables devices, and you getting the data from those devices at every minute, for 200 subjects, over the period of 10 days. Data like this can be explored in different ways, but we will have to go outside of the traditional ARIMA model. One possibility is the so called Dynamic Structural Equation Models. But we may also have very insightful results with simple descriptive analyses of trajectories, like by studying the changes, slopes, peaks, visualizing, etc. When you have many subjects, the goal can be extended from forecasting to comparing groups of subjects, and to some sort of personalized medicine.

Okay, we did not intend to get too deep into this, but just give you a sense of what kind of data can be handled with these models, as well as types of research goals. So, don't hesitate reaching out if you think you can use time series but are not sure how.

052 (14/April/2020) – Measurement Error *to Top*

The journal Statistics in Medicine has published two papers in its Tutorial in Biostatistics section that brings a comprehensive treatment of measurement error. The first paper introduces different types of measurement error and the second paper looks at advanced methods to deal with it. We will talk a little bit about it, but by no means we would be able to cover the full content of the papers in any way that could be called reasonable because it is advanced and long. Instead we will just get you aware of some points we think are more important for our research.

Measurement error is what the name implies: when you collect your data, you are unable to measure things precisely for different reasons. The extent to which your measure is not precise will affect the statistical analysis that you will do, effect that will always be negative since measurement error is a source of uncertainty: the lower the better.

Maybe the most intuitive thing that happens is that you lose power. Let's say that you want to measure subject's age, but for some reason like the subjects not being comfortable telling you their age, you don't get the age precisely correct. This will make it harder for you to detect effects related to age, that is, it will decrease power.

However, I would say that losing only power is the best-case scenario. If subjects tend to understate their age, for example, meaning that what you get is not only an age variable that is not precise in random ways, but an age variable that is biased downwards, then the consequences can go beyond losing power to straight biased estimates and conclusions from statistical analysis. That can be the case even if your measurement error in age is just random (some subjects understate and other overstate their age).

The classical example of a problem caused by measurement error is when you collect age variable that is not precise and use it as a regression predictor. One of the little mentioned regression assumptions is the assumption that the predictors are measured without error, and you only have error in the outcome. The measurement error present in age will cause the coefficients of age in a linear model to be biased downwards, and to the extent that age is associated to other predictors in the model it will also affect the coefficient of the other predictors. This is problematic, even if you can measure your predictor without bias, coefficients of linear models will be biased, and it is hard to say what happens if the model is not linear.

If you want to adjust for the measurement error in age (or in a regression predictor in general), you will need to know something about that measurement error. When measurement error does not exist the reliability is 1, that is, your measure has perfect reliability. So, knowing the reliability of a predictor is one way to know something about the measurement error. If you have such reliability, one way to account for it in regression models is to fit it in the context of structural equation models, where you create a latent variable age which ahs observed age as its only indicator and which will have its variance adjusted by the known reliability of age.

Usually there is little we can do about measurement error and we end up assuming it does not exist, even if sometimes we don't realize that we are making this assumption. But it is an important one, particularly revealing in cases where your measure has poor reliability. It is not uncommon for us to work with data that we know is not reliable.

In the two papers, they go deeper in the theory of measurement error, including by defining different types of errors, different methods of adjusting for it and software. And just to mention here, when you measure age in years instead of measuring the exact age, you get a different type of measurement error called Berkson's error, and in this case linear model coefficients are not biased.