The Podcast *Quantitude*

Greg Hancock & Patrick Curran

Season 4, Episode 8

Missing Data: The New State Of The Art With Craig Enders

Published Tuesday November 8 • 50:39

Patrick 00:04
Welcome my name is Patrick Curran and along with my missing completely at random friend Greg Hancock we make up quantity Dude, we are a podcast dedicated to all things quantitative ranging from the irrelevant to the completely irrelevant. In this week's episode, Greg and I get to explore modern methods for missing data analysis while belaboring quotes from Top Gun with our guest, Craig Enders callsign. The Fly. Craig looks back over the past 20 years of development and missing data analysis to discuss what has worked, what hasn't worked, and what new methods are available now that we didn't have back then. Along the way we also discuss Shawn, not Shawn, going to the movies, grumpy old man mode, wiener boy, Grave Digger, Venice Beach zoom backgrounds, lie awake, hung over juries, Greg's grandmother, shiny objects, Motorola flip phones, Ask Jeeves talking nor walls, mimeograph unscrewing yourself, and who can be whose wing man, we hope you enjoy this week's episode. So I'm not sure if I agree that the pandemic is over. But it sure is fun to start getting back to some kinds of things,

Greg 01:21
given that I just had COVID I'm pretty sure it's not over yet. So much. I'm sorry,

Patrick 01:27
you had not responded to one of my emails. I pinged you on it. And you responded sorry. I've been asleep for 36 hours. It's true. One of the things that I'm enjoying getting back to is our bands are moving back indoors for almost two years. We played in picnic shelters and on basketball courts. But you know what is really fun people who've listened to prior episodes of the podcast know about Sean. Sean is my nemesis because in the quintet Sean plays first trumpet and I play second trumpet. And no matter how much I practice, no matter how much better I get, Sean is always better. And the frustrating thing was Sean is one of the nicest guys you're ever going to meet on the face of the planet. But what is so fun and it's already causing me anxiety. Sean is getting married and the quintet is going to play it as wedding.

Greg 02:27
Oh, he let you be in the quintet. Really?

Patrick 02:29
I am in the quintet. Okay, wow, that's one of my favorite things post pandemic. Nice. What are you enjoying getting back to

Greg 02:37
so one of the things I absolutely love to do with the family, even by myself is I like going to movies. You know, it's very cool that we can stream things at home. But I actually love the physical environment of the movie theater. And for a couple of years, we just didn't go. So I'm finally able to go back into the theater. Take the kids. I am really, really happy about that.

Patrick 02:59
Okay, so I love that as well. I know it's very rare that I do this. Can I go into grumpy old man mode? Would you Okay, so movies you and I went to included Ghostbusters. Bill and Ted's Excellent Adventure. The Matrix Batman Star Wars? Yeah, big blockbuster movies that have been out are Ghostbusters building. Adventure. The Matrix Batman Star Wars. Dude, what are they gonna start writing new material?

Greg 03:32
It's true. And I haven't seen most of those. But there's a part of me that's like, that's like from a different chapter in my life. So I'm not sure

Patrick 03:40
I drove my kid down to Atlanta over the summer, and I laid up in town overnight at a hotel before driving back to Chapel Hill. I went over to the movie Plex and right there on their biggest screen was Top Gun part two, and I thought I Here we go another sequel? Uh huh. It was one of the best movies I've seen in like 20 years.

Greg 04:02
Completely agree. I don't want to give it away for people who haven't seen it yet. But I think they got just about every aspect of that. Right? I want to go see it again. Honestly.

Patrick 04:12
Exactly. It was amazing. If you have any affinity for either action movies, or kind of a romantic tied to the original Top Gun. I highly recommend this now. I just had an idea.

Greg 04:25
I think No good can come from this but okay, what I think

Patrick 04:29
we should give each other call signs for this episode.

Greg 04:34
You say that as though you have a list.

Patrick 04:36
I actually do not because I didn't know we were even going to talk about Top Gun. All right, but I could come up with some suggestions. I'll tell you what, because you know, in reality, you don't pick your own callsign the other pilots give it to you. Oh, I did not know that. So I'm gonna give you three and I'm gonna let you pick which one

Greg 04:55
is okay. Yeah, great

Patrick 04:57
for the rest of the episode.

Insert 05:00
Who are your friends? Pay back, fanboy, what do they call you, Bob. No, your callsign. Bob. Literally

Patrick 05:08
right? wiener boy.

Greg 05:13
It's hard to believe you didn't already have that written down, but okay, go ahead

Patrick 05:16
staphylococcus.

Greg 05:19
Does that actually fit on the helmet? I would just wrap around all the way to the back. I think the staff would hit the caucus

Patrick 05:25
use a small font. Okay, and the third one and I'm drawing from last night on the back deck reading Winston.

Greg 05:32
These are the options. Three wiener boy

Patrick 05:35
stuff. Those are Winston. Dude just picking up this is not brain surgery here. All right, I'll be wiener boy. Okay, wiener boy. Outstanding. Alright, so my choices are Night Fury,

Greg 05:48
wait, wait, wait, wait, wait, wait. I get to come up with choices for you.

Patrick 05:51
I thought you weren't paying attention to that. Oh, hell

Greg 05:53
no, I've been writing them down. Ready? I scribbled down a bunch so I can't really narrow it down to three

Patrick 06:00
clock is ticking. All right.

Greg 06:02
Whack a Mole.

Patrick 06:04
Reaper. Ooh, that's badass.

Greg 06:06
I know. Grave Digger. Not bad.

Patrick 06:08
Okay, these are Monster Truck names. Okay. Why do you not talk about that? What's your third one? That was three actually.

Greg 06:25
So I have a fourth one. Not Sean.

Patrick 06:29
That's it. Alright, so we got wiener boy and not Sean. Done and done. Now. You know what something else that was not phoned in this summer as a sequel was one of my favorite people in the whole world. And I'm not saying that just because he's sitting on zoom right now listening to us and hasn't talked yet. Is Craig Enders. And Craig Enders released the second edition of his book on missing data analysis. And I gotta tell you for a sequel released over the summer this one parallels Top Gun.

Greg 06:58
Honestly, I might even like it a little bit better than Top Gun Maverick. I know. I don't.

Insert 07:04
The end is inevitable Maverick kindness headed for extinction. Maybe so sorry. But not today.

Patrick 07:12
I mean, it's really close. But if you're gonna make me rank order them. Right now, folks, Craig's like Shawn, in that you want to not like him because he's so good at everything that he does, but that he is too nice. He's too funny. He is too engaged in the field. He is cooler than you and me combined. Now

when we got on to zoom about 15 minutes ago, Greg said is that Zoom background? Because it's this like kind of bamboo live plants? Nope. Craig is sitting on his rooftop at Venice Beach and I just can't compete with that.

Greg 07:49
So if you're done talking now, Patrick, I'm going to unmute Craig here. There we go. Welcome, Craig. Hey,

Craig 07:54
thanks. Is this where you hear the sultry sax music?

Patrick 07:57
We can do that.

Tate 07:59
Hi, the Tate Hancock callsign Kenny T. This one's for you Dr. Enders.

Greg 08:12
But if you prefer like some really good trumpet music, I know another guy who's really good. Yep. Thanks

Patrick 08:17
wiener boy not now, not Sean. One thing we love doing when people visit is getting an origin story because every single one of us has some drunken stumbling from one thing to another. That leads us to where we are. Very briefly. Dr. Craig Enders is a professor in the quantitative psychology program at the University of California at Los Angeles. And we're gonna let him introduce himself and give his own drunken stumbling story. So Craig, welcome.

Craig 08:45
Hey, thanks, guys. origin story. So I was a psych major. And, you know, I sort of liked everything about psychology, but then didn't like it, but not to gravitate toward one specific thing. And simultaneously, you know, I had been playing and was playing in a rock band back then my only aspiration in life was to be a musician, right? The only thing I liked to do early was sit around and program synthesizers. So academics was sort of neither here or there for me, it was fine. I was going to college. It was the kid who would like show up late to class and sit in the back and fall asleep. Now where was that? It was at the University of Nebraska and where did you grow up in Lincoln? Same place. So you know, college was whatever it was wasn't an engaged student at all. And you know, sort of stumbling through doing good enough but had no intentions of going to graduate school. Certainly. So then my senior year, I had been putting off this advanced stat class that all the psych majors have to take and sort of dreading it like everybody does, and just had this amazing professor that was life altering. His name's Cal Garban. And you know, the class was this hodgepodge of psychometrics and statistics that a lot of real data sets that were interesting and captivating, and was the first class in college that I was like, wow, this is actually really freaking cool. You know, it was so neat that you could find like patterns and data was just archaeology missions that had all this intrigue. And that was the senior year of college,

graduated, kept playing in the band, we were touring around the Midwest, doing good enough that I didn't have to work or do anything other than just drink on the weekends and get paid.

Patrick 10:36
What was the name of your band?

Craig 10:38
It was called lie awake,

Greg 10:39
lie awake. Yeah. Nice.

Craig 10:41
So you know, doing that. I was the first person in our family to go to college. Right. And this was all thanks to my mother. So simultaneously, I had heard in the background being like, you gotta go to grad school, this band thing isn't gonna last forever, just ceaseless harassment.

Patrick 10:59
That's a good band name right there.

Craig 11:02
two story outhouse. So first of all, I had done nothing to prepare myself to go to graduate school, right? I have a crappy 3.4 GPA or whatever. And literally nothing that I needed to do at a gig the night before I took the GRE and was so hammered that night. I couldn't stay for like the advanced GRE part. I made it through the first half and then just failed because I was still hungover. You guys can cut that out, right? Oh, hell no,

Greg 11:32
it's in the promo.

Craig 11:34
So I called up Cal the guy whose stat class I had, as a senior is like, you know, I really liked your class, what's the next step? And he was like, well, actually, over in the ed psych department, they've got a great group of psychometric folks there, my strategy was, I'm going to just start taking classes. And I'm not going to apply to the program, because I have nothing in my record that would make them be enamored with me. But simultaneously, I didn't have so much competence that I was going to be a different kind of student that I was as an undergrad. So I started taking classes and eventually just got to the point where they're like, Look, you can't take any more classes, you got to apply to the program. And I had done fine. And all the classes at that point, it's sort of turning the ship around, and so went and got my master's, and was dipping the toe in the water and was like, alright, well, I'm gonna go for the PhD. This is fun playing gigs on the weekend making money I can do graduate school during the week. So that's what I did my final year. I didn't know really what I wanted to do after that most of the folks in the program went into industry testing companies and things like that. And there was something about that, that just didn't seem super interesting. So I was like, Well, I'll try a fact if the gig doesn't work

out, I can always hit the eject button and go back into industry. So I got my first job was at the University of Miami. It was there for about four years. And then my advisor, Nebraska, Debbie Bandoleros, left. So I went back to Nebraska for a couple of years, then was fortunate enough to get a gig at Arizona State where I was at for 11 years working with a bunch of my heroes from graduate school, right. And that was just the thrill of a lifetime. And so then I ventured out here for a summer and rented an Airbnb and Venice Beach. And it was like this is the coolest place. UCLA happened to have a open rank position and was fortunate enough to get an offer here and decided to come hang out by the beach.

Patrick 13:47
So just another standard. Nebraska boy, rock star grads, formulaic way to become a quantitative methodologist. Yeah,

Greg 13:58
I will say I remember you when you were a spiky gelled up hair grad student working with Debbie. And we were presenting in the same session at the Ara conference in Seattle. And because it was in Seattle, my grandmother was in the audience. And she was 84 years old. At the time, I had gotten up giving my talk, I had acknowledged that my grandmother was in the audience and you had the next presentation. And you got up and you said, God, I hate following Greg and his grandmother is in the army. The pressure is insane. And it was so cute that you would acknowledge my grandmother, so I will

Craig 14:35
never forget that was insane pressure.

Patrick 14:37
Now one of my early recollections. We were out at a bar somewhere, but there were three or four or five of us out and you had ordered like a whiskey neat. And you looked in and there was a fly in your drink. And you were so polite, and you called the waitress over and you said I am so sorry, but there's this fly in my drink. Could I trouble you for another one, and the woman was mortified. And she was like, I'm so sorry, I'm so sorry. And she went off, but it was really busy. And we're sitting there in the drinks on the table, and we're sitting there and the drinks on the table, and you pick it up and you start kind of angling it and taking little sips kind of around the fly. You dangle it and flick it with your finger and take another sip. So that's my first memory of you, as you were perfectly fine just to drink around the fly.

Greg 15:32
So it is at this point in the conversation that I feel that we need to have Craig have a callsign Oh, and so far, I have lie awake, obviously in reference to his prior life. Yeah, whiskey neat. Not bad. Oh, I like that. And I don't know what to do with fly. Just say the fly the fly. There we go. So the flight so your options are lie awake, whiskey neat, and the fly,

Craig 15:57
think you gotta go with the fly.

Greg 16:00
Or you go winter boy, not Sean and the flow. All right, at some point, we should probably talk about missing data. Patrick, you winna get us started? Talk to me, goose? No, no.

Patrick 16:12
I was at a conference, one of the ones that Penn stayed on the new methods for the analysis of change. And I think I was like a second year professor, I had been very fortunate to be invited to present. But John Graham and Joe Schaffer are in the audience, Linda Collins, all of these people, and I get up and I have this sample of data I was using, and I did list wise deletion. And I still remember as I stood up on the stage, and I paused and I said, Hello, my name is Patrick, and I'm a list-wise deleter. And everybody on cue went welcome, Patrick. Phenomenal. So one thing I find interesting is with your book coming out in 2020, to 2002, was the very impactful Schaefer and gram view of the state of the art of missing data. And I love round numbers. I like even things I like flat things,

Greg 17:18
and shiny objects.

Patrick 17:21
And a 20 year updating on what we see as state of the art, I think, is a really nice balance with your second edition of your book coming out, what is the new state of the art because if you think about what was state of the art in 2002, I was thinking about this a little bit last night, the iPhone was introduced in 2007. Wow, I looked up what was the top selling phone in 2002, there was a Nokia and then a Motorola flip phone. Now, through this special TV offer, you can receive a Motorola flip phone with cellular service for just pennies a day. In a way now, it's not undermining Shaffer, and Graham at all. That was a profoundly important paper, but it's kind of the Motorola flip phone of our view of missing data. So why don't you be the iPhone 14? What is the new state of the art in missing data? Well,

Craig 18:25
first, let me defend the flip phone. You can still buy those and they're fantastic devices. That paper is most highly cited paper. And it's like methods by a longshot. And I think it's still really like a must read for anybody who wants to get into missing data stuff. So much of what they talk about in there is still sort of our daily driver methods that we use right simile, or multiple imputation. Those are still solid go to methods. So yeah, I'll tell you what I'm enamored with. And that I think, is the new state of the art. It actually dates back to that same period, late 90s, early 2000s, there was a group of papers by this guy, Joseph Abraham and a couple of his colleagues, and they sat dormant for 15 1617 years, and he laid out a different way to approach missing data handling. So you know, most of the methods that were predominant at that time are centered around applying a multivariate distribution to missing data, right. So multivariate normal being the typical one. And Abraham's approach, let's call it factored regression specification. So his idea was to take this joint distribution, this multivariate distribution and factorize it into the product of a bunch of univariate distributions. So in the simplest case, let's say you've got x and y, and we could model that with a bivariate normal distribution. So his approach is to break that up. into a product where it's y conditional on x, just a regression model multiplied by just a marginal distribution for X. So if we had three variables, let's say y x 1x, two, we'd have y given x one and x two, then we'd have x one given x two, and then a distribution for X two. And so the really cool thing about that

factorization approach is each of those regression models can be of a different ilk, essentially, right, we could have one of them via linear regression and other be a logistic regression. One could be account model, it could be anything. So his approach was to say there is a joint distribution, but we don't have to model it, we don't have to say what it is, we're going to just model the individual factors on the right side of the expression, each of which is a univariate regression model. So that opens the door to all sorts of things that we can't do well with, say a multivariate normal models to stuff like interaction effects with incomplete data, you know, those are very much at odds with a normal distribution, or random slopes and a multi level model are curvilinear effects, or mixtures of categorical and continuous things, those individual models could be really just about anything, they come together into a joint distribution that you sort of acknowledge exists, but you never have to model it formally or explicitly.

Greg 21:33
May I ask two questions? Yeah. Fireworks. One is, can you stop the birds from singing behind you? Because you're writing a damn Disney cartoon? Okay, thank you thing, too, is you describe the factorization. And you did it in a particular order. Is there any role that the order of the factorization plays in this,

Craig 21:52
the key I guess, is that you want one of the terms to correspond to your vocal analysis, right. So like y given x one and x two, that's my vocal model. And then the rest of the models we could think of as essentially nuisance models a way to deal with, you know, in a regression model, incomplete covariates that might have different metrics, or maybe be non linearly related to one another. So yeah, there's different orderings. But it makes sense to go from least amount of missing data first, like make those last pieces in the factorization, because if a variable is complete, you could drop that model completely. A dysfunctions is the constant essentially, right. And so it would be one less model that you have to estimate in this chain of stuff that you're dealing with, and maybe categorical variables first, ordered by sort of least to most amount of missing data followed by continuous variables in the factorization. Those are, I think, more practical recommendations, I guess, just to place

Patrick 22:52
it in a broader context, would that still be of use when you have complete data? Because if you have this very complicated multivariate distribution that you can't express in some closed form, but can build it with a combination of these individual ones? It seems like that could be beneficial even when you have complete data. I think

Craig 23:13
it depends on the model for a regression model. I don't think that there's any benefit to that. But I think, you know, applying that logic to a structural equation model with complete data allows you to do some really interesting things. Because that framework, the way that latent variables are viewed are just as missing data. So the latent variables essentially get imputed along the way. And that factorization applies, like latent variable model allows you to do all sorts of things like an interaction between a latent variable and a nominal variable is the easiest thing in the world to do in that framework, you know, or some of the stuff that you work on Patrick, with the nonlinear factor models with those complicated constraints, you can do that really easily in that framework without any of the constraints just let the

grouping variable or whatever the background variable is interact with the latent variable. And so I think are latent variable models that raises some really interesting avenues that we don't necessarily have at our disposal when we're attacking things from sort of a normal distribution perspective.

Patrick 24:19
FIML was the default in everything at this point. And MI you have to work harder for but it's still pretty broadly available. How would you approach a factored regression approach

Craig 24:30
the last six or seven years of head grants from IES and we built we being myself and former grad student brand Keller have built a software package blimp that started off as sort of a modest multi level imputation program and the first grant and the second grant was really to expand it to allow for missing not at random models and those models require two equations. You've got your vocal regression. Let's say some Multi Level model and then you've got the missingness model. And we sort of realized along the way that if we can estimate two equations in this framework, we can estimate 50 equations, right. So it sort of went from barely basic start to just sort of blowing it up into a full multi level latent variable modeling package that handles all sorts of different types of variables, categorical continuous count variables. So our goal was to put this powerful, flexible stuff into a package that anybody can open up in the US without deep level knowledge about what was going on under the hood in so my website applied missing data.com/blimp got an app up there for Windows and Mac, and Linux does all sorts of latent variable multi level modeling stuff all within this factored regression specification approach.

Greg 25:52
So I got on Netscape and when to Ask Jeeves to try to find your website, but I couldn't. Could you tell me again what it was

Craig 26:01
WWW.appliedmissingdata.com.

Greg 26:03
Okay, maybe I should upgrade my 2002 technology. So you mentioned blimp, I'd like to try to make a connection between the software blimp and the second edition of your book that just came out this summer. That was better than Top Gun, the second edition of applied missing data analysis. What I will say just to start off, though, is a lot of second editions really are like a lot of sequels to movies, where there's some updates to References, maybe some new graphics, and all of that. And tada, it's a second edition. And part of that is fed by publishers who are always pushing you to get out a fresh edition. But I will say you took this top to bottom and redid it. And I had the pleasure of already going through this book. Can you tell us the role that blimp plays in the second edition? And then just more broadly, about the second edition?

Craig 26:51
Yeah, you're right. It was a complete overhaul, I literally deleted everything and started over part of that, you know, as you guys know, from teaching many, many workshops, you just find a better way to explain things over time, right. And so it felt like I could do a much better job explaining this stuff 10

years later. So that was part of the impetus. You know, and certainly blimp plays a central role, though the book is software agnostic, I don't really mention any programs by name in the book. But certainly it plays a big role throughout all of the base stuff and all of the multiple imputation stuff. That's what I was using for all of the examples. But I think the thread that goes across all of the chapters is the factored regression piece. So those models are available for demo as well. Actually, in much sort of simpler rudimentary form, I would say at this point, m plus does a few things that fall into that factored regression heading, there's a package in R MDMB that does factored regressions for really just regression models. So there's some backward regression stuff on the thermal side. But that I think, is sort of the thread that goes throughout the book is tying in all the new developments in that space. And so part of just starting from scratch was I want to get rid of some of these cringe worthy explanations that I gave in 2010. But then

Greg 28:17
the ones that I use in my classes right now is that those ones Awesome, thanks,

Craig 28:21
unicorn and the talking narwhal.

Greg 28:26
Very nice.

Craig 28:28
But you know, as I started to sit down, I pulled up all the Word documents from the old chapters. And I just became really apparent quickly that to shoehorn in all of this new stuff, it would feel like really disjointed and a bit nonlinear maybe. And so it just seemed like a better idea to hit the delete button and start over.

Greg 28:46
Wow, I experience a very small version of this just every year when I go to teach a class that I've taught before. In my head, I might have evolved and matured in the way that I want to explain things are the materials that are crafted again in my head that I could use, but it's so easy to go grab the folder, grab the PowerPoints, or whatever it is. So I really think that you should be commended on that that was a brave thing to do.

Patrick 29:11
And not only that is it's one thing to do away with the talking narwhal, which now that's all I'm going to be able to see for the rest of the afternoon. But Craig, let me ask you this. Was that really just how you wanted to tell the story? Or do we as a field need to think about a rewrite on how we think about missing data? So one thing that drives me insane when I review articles is someone will say we use full information, maximum likelihood and thus missing data or not a problem. Yeah, and I have a rubber stamp in my drawer that I pull out and have a brief paragraph about how it's under assumptions and that the authors need to describe the missing data and things like that. But is it hyperbolic to say that we have come a remark probably long way from mean imputation and last value forward to endless

strides deletion to PHYML and multiple imputation. But do we need as a field to think about updating how we approach missing data? Yeah, great

Craig 30:15
question, I still find great value in the flip phone, actually, the way that I was seeing this is we've got these, let's call them classic approaches based on a multivariate distribution. And those things are still insanely useful and easy to use. And we know a lot about them. And we've got these really sexy, new approaches that let us model combinations of weird variable types that don't plug into a multivariate normal model. But you know, those old methods still don't behave that much differently than the newer sexier ones, in a lot of cases. And so I don't think it's a situation where we should abandon these classic methods, I think they're still extraordinarily useful and do surprisingly well, in a lot of cases where you wouldn't think that they would do well. So I sort of see these things as complementary tools in our tool belt, rather than, you know, let's wipe the slate clean and do away with these older approaches.

Greg 31:17
So if you get that manuscript that says that it used PHYML, you're not pulling out the rubber stamp necessarily to say that time has passed, but you just want to know that it's used under appropriate conditions.

Craig 31:28
Yeah, exactly. I think there's things that thermal does really well. And then there's things that FIML does pretty terribly where you're going to expect to see some bias. So yeah, I agree with that.

Patrick 31:40
What are those situations where it does terribly,

Craig 31:43
you know, I think anything with a non linearity in it and interaction term, or curvilinear effect, or a multi level model, perhaps with incomplete covariates, it's really the covariates that are the problem, the predictors, that things on the right side of the equation, when those go incomplete, that's where you have to be really careful, if you've got an incomplete predictor, let's say it's exerting a random slope and a multi level model, or let's say that it's part of an interaction of fact, when you go to impute that variable, B, it them all sort of implicitly, you know, imputes, the data or Bayes or whatever that variables distribution is heteroskedastic. And so you have to be able to model that feature of data. And so a lot of software packages will gladly give you an answer when you use them all. But we know from analytic work, and also computer simulations, that my results might not be so great on the back end. So that would

Greg 32:41
put you in the danger zone.

Craig 32:49
Correct. You're good? Yeah. Boy.

Greg 32:52
Did I take your breath away, Patrick? All right,

Patrick 32:59
I'm going to edit this episode. We're gonna be gone negative Ghostrider. The pattern is both. Okay, I was gonna say one more, you know, what I'm gonna do just to get in your head for the rest of the

Greg 33:11
day? No, don't do it. Look, travel, travel

Patrick 33:19
by. But just to clarify is my rubber stamp is not that you shouldn't use PHYML on their more modern methods. My rubber stamp is PHYML doesn't mean missing data or not a problem, right? These are routine kind of things that you see in the literature, I will read papers that don't even say what percent of data are missing? Or what patterns of missingness exist given known covariates. What I want to see is a paragraph that describes all of those things of the total sample, what percent were complete? What percent were missing at different time points, how do those relate to observe covariates and then transition from there into given the characteristics of our data, we believe full information, maximum likelihood is an appropriate method, you know, whatever. That's more My thing is that there's a whole generation of students that seem to think we don't have to worry about missing data anymore. Because so and so package can handle it.

Craig 34:24
Yeah, absolutely. It's demoralizing how little the needle moves over time, right? How still terrible the reporting practices are.

Patrick 34:34
How would you improve it? If you said as one of the leading experts in this field in the world, here are X number of things? I think an applied researcher should report as part of their manuscript. What would those be? We'll go read chapter 11.

Craig 34:49
No, I mean, all the things that you said I agree I have in that chapter, sort of a laundry list. You know, many people have complained about the state of report reading practices and many people have catalogued things that they would like to see reported. And I just really stole all of those and put them in one little package in chapter 11. And we've got online supplements now, right? We don't have to cram this into two sentences in the methods, we've got infinite online space, certainly, as you said, talking about, say, how much missing data that they have them exploring our assumptions about why the data are missing, right? It's easier than ever to apply missing, not at random models to data, those aren't out of our reach anymore. They're simple to use, and not appropriate for every situation. But certainly some are doing sensitivity analyses that examine the impact of our assumptions about missing data. You guys both expressed your love for demo in an earlier episode, I think one of the things that makes imputation or Bayes really appealing to me is that we can see the assumptions that we're applying

when we do FIML. Right, we can plot the imputed data and get a visual bead on what we're assuming about the distributions of the variables. And it's kind of shocking sometimes to see what the missing data models imposing on your results. So certainly doing sensitivity analyses around distributions that we apply to missing data, I could go on and on and on and on. Right, and probably that's overkill. But certainly I page with a little more detail. And an online supplement is not too much to ask.

Greg 36:33
I have chapter 11 in front of me. And it's wonderful. It's a wrap up to the whole book. And it helps people to know about reporting, just like we're talking about. Let me just read the headers for the recommendations. And Craig has been drilling down a little bit into those but I highly recommend that folks take a look at this recommendation one talking about your missing data rates, recommendation two talking about distributional assumptions, recommendation three, the missing data process recommendation for any auxiliary variables that you might have recommendation five, missing data handling methods, recommendation six, the software tools and implementation details, recommendation seven sensitivity analysis, and recommendation eight, Bayesian estimation and multiple imputation. So throughout, Craig has a very nice flowchart of ways that you can think about things that you've done and ways to make good decisions and to report about those decisions. So chapter 11 is a beautiful wrap up to a fantastic book,

Patrick 37:33
you intrigued me a moment ago, you had an almost throwaway line about Well, now we can handle MNAR and I perked up a little bit there because when I teach this, I talked about MCAR and MAR. And I always have a bit of a shrug and say, if you're MNAR I think the technical term I use is you're kind of screwed. Maybe I need to update my own teaching. Now granted, I'm still using curl than yellow transparency slides so that I had from like 1988, some dos, here's my ditto and quit sniffing them. Everybody put them down, you don't need to sniff the mimeographs. Tell us a little bit more about the MNAR

Craig 38:19
Yeah, sure. So just as a reminder, you guys did a really amazing job of describing missing data mechanisms and missing data processes in one of your earlier podcasts. And just to sort of reiterate what those are because the terminology is awful, right? So missing completely at random is a situation where think about the causes of missing data, those are really unrelated to the variables in our model. And our default assumption most of the time is missing at random. Or maybe it's better to call it conditionally missing at random. So it's the idea that once we condition on the observed variables in our model that missing this is purely haphazard after that. So another way of saying it is that the unseen score values that we don't have access to those don't play a role at all. There's no tell us anything about why people might be missing above and beyond what we already know from the observed data. And so that's the stock assumption that we make with demo or days or multiple imputation. And then the third one missing, not at random. This is where the unseen score values themselves carry information about whether a person is going to have missing data or not. So it's my level of substance use that influences my decision to not report that to you.

Greg 39:48
So how do we deal with that last one in practice the missing not at random, right? What are some ways that we might approach that?

Craig 39:54
Yeah, the two major frameworks that we could work from our selection modeling and Pattern mixture models, selection models, essentially, you've got your vocal model, whatever that happens to be. And let's say you've got, Patrick, in your context, this is probably a huge thing, when you're talking about substance use and stuff like that, it's reasonable to assume, you know, they will might not report their usage patterns because of their usage patterns. And so, you know, we've got this dependent variable that maybe isn't conditionally missing at random. And so we would pair that focal model up with a second regression model where the missing data indicator is the dependent variable. And we're predicting that missing data indicator from the very least the dependent variable itself. So why is predicting it's missing us, but then there could be auxiliary variables in that missingness model, some of the covariates from the focal model could be in that missingness model. And so simultaneously estimating our focal model, and that missing this model, if we get the missing this model, approximately correct, we'll adjust our focal model estimates in a way that removes non response bias can think about those models as like a mediation process, almost like x influences y and then y influences missingness. And maybe some of the x's also influence missing us directly. And then pattern mixture models flip the script and are more like a moderation process where that missing data indicator becomes a predictor variable. And essentially, we're saying that subgroups of people with and without data on our dependent variable potentially have different parameter estimates, they could have a different intercept or a mean level on the dependent variable. Maybe the effect of an intervention works differently, depending on whether you have data or not on the dependent variable. So those are the two major camps. And there's lots of other flavors of those models, sort of variations on that theme. But I think those models are fairly straightforward, you have to be very careful with them. I agree with your sentiment that you have this, you're not entirely screwed, but you got to be really careful unscrewing yourself. Yeah, so I think, you know, we know a lot about these models and know how they behave on real datasets, they don't feel quite as mystical maybe as they did 10 years ago.

Greg 42:29
So I love all the missing not at random stuff that you're talking about. And what it makes me think of is, what are some of the really cool areas where people could be doing research, we have a lot of people out there who are listening who might be toward the end of the graduate program, they're thinking about areas for dissertation topics. And this feels like,

Insert 42:48
this is what I call a target rich environment.

Greg 42:52
So can you tell us what some of the areas are that you think are good areas for people to be working in?

Craig 42:57
Yeah, well, I think there's a lot of meat left on the bone with these factored regression specifications. We were building the software package, they sort of have this realization, oh, we could do this, we can do this, we can add this variable in here, this count variable, or this skewed variable, there's infinite combinations of models that you can pack into these factored regression specifications. And the theory sort of says they should all work and the software runs the models just fine. And you get results that seem to make sense yet. Also, it's ripe for computer simulations, that is to try to understand when these models work, and when they break down when they reach their optimal performance. So many of the studies I think, focus on models with continuous outcome Zahn simple interaction models with two covariates, interacting, who's got all sorts of discrete variables we can throw into the mix now and different types of nonlinearities. And really, the sky is the limit. I think there's a whole lot of stuff to be done there. You know, and certainly also in the multi level context is well, right. So these factored regression specifications we've been talking about, we can use them for two and three level models with all sorts of interesting stuff going on with the covariates can be different metrics and exert different types of effects on the dependent variable. And there's been a lot of research on multi level models with factor in specifications. But again, it's pretty simple models with continuous outcomes. It's really just scratching the surface what these models are capable of. And so I think for people looking for dissertation topics, there's plenty of room to dig a little bit deeper and try to understand these models so we can apply them more carefully. Something

Patrick 44:53
you said earlier that I really resonated to was sensitivity analysis. Yeah. Because your exactly right as a lot of work needs to be done in terms of when do they perform well, when do they not perform well, but I think some thoughtful work could also be done on how to use them and drawing up some general heuristics that involve some of these methods of sensitivity analysis, because I think a lot of us who do substantive research are deeply dedicated to doing the analyses correctly. But at the end of the day, we want to know, Is something going to change my discussion section? Yep. And so if you do a straight up PHYML, and get your results, can we incorporate some of these new methods to increase our confidence in that in terms of a sensitivity analysis, but if it does lead to a difference, in conclusion, then that's an intellectual goose for us to try to better understand why the two methods differ. So I'm very excited about new projects that lay out some general heuristics of how you would go about using these methods in practice.

Craig 46:03
Yeah, agree 100%. And I think another avenue for future research, certainly, for folks who are in the SEM space is a how to assess fit in these models, I think is a real sort of tricky thing, right? When you've got categorical covariates interacting with a latent variable. And so all of our usual fit tools really kind of go by the wayside, there is not a reproduced covariance matrix that we could pit against the sample data in the same way that we do with conventional sem techniques. And there's been some work on this, like Ed Merkel has worked on this. And Bolin has done some things other folks have done stuff. But it's really interesting to sitting down and fitting sem models with these factored specifications. Because stuff we sort of take for granted when we're modeling and multivariate normal framework, even how you identify the types of constraints you put on a model to identify it can have a surprisingly big impact on the convergence of the model and things like that. So there's a lot of interesting little stuff

to figure out about these models and practical things for where the rubber hits the road, applying these things to real data.

Greg 47:13
I really love listening to you describe what some of the new areas are that you've been working on that you've put into your book, and then areas that are right for people to do work in. As always, your explanations are just spectacular. So one thing that I've known about you all the way back to your spiky gelled up hair days, is that you are a wonderful explainer a wonderful teacher. For anybody out there who hasn't had the pleasure of taking a workshop with Craig and I actually have taken a workshop with Craig, I highly recommend that you do it. He's just a very, very gifted teacher.

Patrick 47:48
I completely agree. Craig was foolish enough to work with Dan Bauer and me and has a workshop with center stat. I have to admit, Craig, when you first taught I was working and listening to your class, because I was learning these things a new myself. I second, Greg's characterization is I think you're one of the clearest teachers I've ever had. And so anyone who has access to that, I would highly, highly recommend that. Well, thanks

Craig 48:17
for that both of you guys. I mean, in truth, both of you have really informed how I teach. I was sitting in the back of that conference with Greg's grandma taking notes. You guys are both master presenters and master teachers and I've soaked in your expertise over the years and borrowed and stolen things.

Patrick 48:37
I feel a hug coming on, which means I think we need to wrap this up.

Craig 48:40
Yep.

Patrick 48:40
Thank you, Craig.

Greg 48:41
Yeah, so definitely Craig, you can be our wing man anytime.

Patrick 48:46
Bullsh*t. You can be mine. Thank you, everyone.

Craig 48:54
Thank you guys.

Greg 48:55
Take care everybody. Thanks so much for joining us. Don't forget to tell your friends to subscribe to us on Apple podcasts, Spotify, or wherever they go to fill the holes for what's missing in their lives, not at

random. You can also follow us on Twitter where we are at quantity food pod and visit our website quantity and pod.org where you can leave us a message find organized playlists and show notes. Listen to past episodes and other fun stuff. And finally, you can get cool quantities merch like shirts, mugs, stickers, and spiral notebooks from redbubble.com where All proceeds go to donorschoose.org. To help low income schools, you've been listening to quantity food, your home for all manner of non normality. Today's episode has been sponsored by quantitatively relevant Top Gun movie quotes. Number one, how you feel when you try to run a Monte Carlo simulation study of factor models with sparse discrete items using numerical integration on your laptop. Feel the need number two, how you fared on that grant submission as a brand new assistant professor

Insert 49:59
crashed and burned On the first one it wasn't pretty.

Greg 50:02
Number three how you're feeling when all those projects you thought it'd be fun to say yes to all need you to do stuff for them in the same week. Your ego is writing checks your body can't cash. And finally number four, how we hope quantity has made you feel about MANOVA and post hoc power analysis. You have lost love and is this NPR? That is most definitely a negative Ghostrider.