Hi, everyone.
Welcome to the departmental seminar of the Departmental Biostatistics, Yale University.
I'm pleased to introduce you Linglong Kong. He was associate professor of the Department of Mathematical and Statistical Sciences at the University of Alberta.
He's research interests are on, and correct me if I'm wrong, functional and neuro imaging data analysis, statistical machine learning, and robust statistics and quantile regression. So today, he is gonna talk about his work on general framework for quantile estimation with incomplete data.
Thank you, Linglong. And whenever you're ready.
Thank you Laura for the introduction. And also thanks Professor John for the invitation.
I'm very happy to be here, although it's way too early.
So today I'm going to talk about general framework for quantile estimation with incomplete data. So, this is a joint work with Peisong from University of Michigan and Jiwei from University of Wisconsin-Madison, and Xingeai. And we started this work when at the second year when I started my position at the University of Alberta.
I know Peisong a long time ago before he was a student, and at that time he just started his position as assistant professor at the University of Waterloo. And I invited him to visit me and afterwards, he invited me to visit him. And we feel like we visited each other already, we should get something done. But I remember that I’ve known where he stayed in his office at the University of Waterloo and thinking about what do we have to do together. And eventually we thought, "Okay, what I’m good at and while all my research area is quantile regression.

One of the research area of Peisong is missing the data." So we said maybe we can put them together, then we are write a couple of formula on the paper. Then we feel like, "Okay, we get a copy already.” Then we went to have a dinner. And then one year later Peisong send me like two pages to trap, said maybe we should continue it. And that’s the first scenario in this topic, I’m gonna talk about.

I sent him my feedback. And then another half year, I sent him my feedback. I said, "Why don’t we make it more general,
So this semester we’re going to be able to apply to honor other scenarios. And then we both feel it’s good idea, then we started working on it. At that time, Jiwei was posed to at a University of Waterloo and Xingeai where my post are. So, we thought together and started a project. Eventually, I wound a project that I’m kind of proud of.

So, what’s missing data? The missing data arise in almost all serious statistical analysis. Missing on values are representative of the messiness of real world. Why we would have missing a missing value, it could be all kinds of reason. For example, it may be due to social or natural process. Like for example, a student get a graduate, get a job out in clinical trial, people get died, and so on. And also could happen that you survey. For example, in certain question asked, only asked respondent answer yes, to continue to answer certain questions. Or maybe it’s the intention missing as a part of a data collection process. Or some other scenario including random data collection issues respondent refusal or non-response.
So, mathematically how we categorize these kind of missing, and here is the three scenario.

Now, first scenario we call it missing completely at random. What does that mean?

That means the missingness is nothing to do with the person being studied.

They’re just completely got missing, it’s nothing related to any feature of this person.

The second scenario is missing at random. The missingness maybe predictive from some auxiliary verbals auxiliary information.

The third one is a very hard one, is missing not at random.

The missingness depends on observed the information and sometime even the response itself.

So, the missingness is specifically related to what is missing.

For example, a person to not attend a drug test because the person took drugs the night before.

And therefore the second day, he couldn’t make to the drug test.

Couldn’t get to that drug test result. These are three missing mechanism.

How do we handle those missing data?
There are many strategies. For example, the first one would be, well, let’s try to get the meeting data. That would be great. But in reality, that’s usually impossible. But the second is, well, as we have incomplete cases, let’s just discard. Just analyze the complete case, right? But these could cause some other problems. We will talk about it. And the third one is we replace missing data by some conservative estimation. For example, using sample mean, sample median, and so on. The first one is we are trying to estimate the missing data from other data on the person. We use on sort of more sophisticated method to impute. Now in particular, mathematically speaking, the strategy we are using today do to deal with missing data, the first one is a complete case analysis. These are very simple, okay? We just analyze compete case, okay? And we only analyze in consideration that individuals with no missing data. Sometimes it can provide good result, but the estimation obtained from this complete case analysis
maybe biased if they excluded individuals are systematically different from those included. So hence, if the complete case would be a good representation of those missing case, then this method would it be fine. Otherwise, if the complete case is quite different from those we miss, then all result can be biased. And then there’s inverse probability weighting method IPW. This is a commonly use method to correct the bias from a complete case analysis. What does that mean? It means, okay, we give each complete case a weight. This weight is the inverse of the probability of being a complete case. Well, this can also cause some bias if this IPW relies on the data distribution. The first strategy is more sophisticated to do these multiple imputation. It’s quite common method, especially nowadays in genetic study. How do we do multiple imputation? We create multiple sets of imputation for the missing values, using imputation process with a random component. Now, we have an full data set. Then we analyze each data set.
Those full data set can be a little bit different. Can be slightly different because the randomness of the imputation process. Anyway, analyze those data set, complete the data set, and then we get all set of parameter estimates. Then we can combine those result. We can combine this result, and we hopefully we get a better result. The multiple imputation sometimes works quite well, but only if the missing data can be ignored. And also, we have a good imputation models. And while it depends on the nature of the data, the auto mind depends on what kind of imputation model we are going to use. Now, that’s how we deal with missing data, the strategy we happen to use to deal with missing data. But let’s matched them together in terms of missing data. How we use these meeting dates age to deal with different missing mechanism. For example, if the data is missing complete at random, now in this case, the complete case analysis is quite good. Multiple imputation or any other imputation methods is also okay.
Is also valid.

So, this missing complete at random is the easiest case to deal with.

What if data is missing at random?

Then in this case, some complete case analysis are valid and multiple imputation nearly is okay too, if the bias is negligible.

Now in a certain case, if the data is missing not at random, then we have to model the missingness explicitly.

We need jointly modeling the response, and also the missingness.

In practice of course, we try to assume missing and random whenever it’s possible and try to avoid to deal with missing not at random situation.

Sometime we have data always missing not either random. Think in that case center or there is one special issue dedicated to missing data, not at a random situation.

Now, we have different strategies.

And that they state different strategies have different advantage and disadvantage.

For example, multiple imputation is generally more efficient
than IPW, but it's more complex.

And the imputation and IPW approach require to model the data distribution and the missingness probability, respectively.

Imputation, we need to model data distribution.

IPW, we need model the missingness probability.

And also, for all kinds of strategy, we would have have good property, only if the corresponding model is correctly specified.

Most existing method are vulnerable to these model misspecifications.

Of course can use nonparametric method to reduce the risk of model misspecification, but it’s often impractical due to the curse of dimensionality.

So now, how do we deal with this model misspecification?

We have some method available. For example, we can use a double robust method.

In particular, in double robust method, we have this augmented IPW.

We are not only model the missingness probability, but also the distribution.

Why is double robust? Because the result would be confusing if the model is correct.

If the way we model missingness probability
or the way we model the distribution is correct, then we would get consistent result. And that’s why it’s called double robust.

Well, now that we are not satisfied with double robust, what about we can a multiple guarantee? So, we have these multiple robust.

This is a proposal by Peisong. And they multiple robust method is proposed to account for multiple models for missingness probability and the distribution. In double robust, we can only one model for missingness probability and one model for data distribution. The good thing is the estimation result will be consistent if either one or the model is correct.

So, we are looking at missing at random. We assume on the observed data are ID. So we have data R, R Y XT. We use it to missingness, and the IPW estimator essentially we are trying to solve these equation. And here, these is the probability,
although makes complete case. And IPW is consistent, only if this \( X \) is correctly specified. And then, then from the equation, we can get consistent estimate of those permit we are interested in. This IPW. The other one is imputation. For imputation, we need model that take distribution. And here we have on the model of a \( f(Y|X) \) And as you can see, we have our imputation for those missing data. This imputation is consistent, only if this state distribution is correctly modeled, this \( f(Y|X) \) is correctly modeled. Now for these augmented inverse probability waited method, we actually combined these two together. We had the first part from IPW, second part from implication. So the estimation result would be consistent if either this model for missingness probability or the model for data distribution is correctly specified. Well, for multiple robust method, they have a serious model for missingness probability and a serious model for data distribution. And all result would be consistent, if any one model is correctly specified. Well, this is something
I just get a quick review about this missing data. Like I said, this is the part Peisong is one of the Peisong research area. For me, my research area is quantile regression. So, internal quantile regression at that time we were thinking, "Okay, those methods, these IPW, AIPW or double robust method, multiple robust method, had been quite well studied for when we model the conditional mean. Therefore, condition of quantile, there are not a lot of methods available. Why we care about the quantile? A quantile not only provide a central feature of the distribution, but also care about the tail behavior. And also under very mild conditions, the quantile function can uniquely determine the underlying distribution. So, there are a lot of advantages to model the quantiles. Then, we decided to study these missingness in quantile estimation. In particular, we proposed a general framework for quantile estimation with missing data. So, our proposed model, these kind of framework, can do a lot of estimation for missingness in quantile estimation. But in this paper,
we particularly applied all proposed method, these three scenario. Okay, three commonly encountered situation. The first one we trying to estimate the marginal quantile of response. This response get some missingness. Well, there are fully observed covariates. That’s the first scenario, response gets some missingness while the corresponding covariates get fully observed. The second scenario, we are looking at the conditional quantile of a fully observed response. So, we have some missingness for covariates. And then the third scenario, we are still looking at the conditional quantile of a response. And in this case, the response gets some missingness and we have fully observed covariates and also extra auxiliary variable. Now, let’s look at the first situation. We want to estimate the marginal quantile. In this scenario, we have the response gets some missingness and we have the covariates fully observed. Now, let m to be the number of subjects with data completely observed.
Then our method consists of the following five steps. The first step, we calculate this or estimate to this. This isn’t related to the missingness probability, okay? The way we estimate this, is by maximizing the binomial likelihood. So, the first step we estimate the and then we get estimate of the missingness probability.

Okay? The second step, we calculate gamma. This gamma is related to this data distribution. This gamma is a parameter related to the distribution.

And then the third step is we can sort of preliminary estimate of the quantile or the marginal quantile through these imputation process, by solving this equation. And as you can see this is quite close to the AIPW scenario. Okay? And in this equation, this five is the score function of quantile lost function. This prosaic is $r - i(r < 0)$. This is the generalized derivative of quantile lost function, okay? Here, this one can not be exact zero.
The reason this phosaica is a non-smooth function.

and it sometime it won’t be exact here.

Basically the first step, okay?

Now, we have a preliminary estimator of the marginal quantile.

The first step is the case that of method

is where the multiple robustness is coming from.

Now, we calculates weights for the complete case.

In total, do we have m complete case.

For each case, we calculate the weight.

As you can see, the weight is determined by three parts.

The first part is related to this alpha,

which is related to the missing probability, okay?

Missing probability.

The second part is related to this gamma.

This is related to the data distribution.

The third part is related to this cube.

This preliminary estimate of these marginal quantile,

which is related to this self step.

As you can see from the first three step,

we are trying to get ready for this,

to get the estimate for the weight for the complete case,

for this complete case.

And also, we have our parameter,

though is obtained through
minimizing these equation, through minimizing this equation.

Now, after we calculate the weight we get off final estimate of our multiple robust estimate by solving the following with estimated equation.

This wi is the width. We estimate it from the first four steps.

And this posy is a score function of quantile loss, okay?

Now, you may get wondering on what’s going on

with these five steps.

And let me try to explain it one by one, okay?

In the first step, we get the estimate of alpha, okay?

In sense trying to model they missingness probability, okay?

Now, in the other case, the missingness probability will not be correctly specified.

Well, in the second step, we only estimate gamma.

We are trying to model the data distribution.
and we have models for the data distribution. And then the third step, we are sort of doing some imputation as made of these marginal quantile. And these marginal quantile will be correctly estimated, if those data distribution is correctly specified. The step four is typical formulation of an empirical likelihood program. I will getting back to this in the next slide, why it’s a empirical likelihood program. And this is a key contribution of methodology. Now, in step five, we have the structure of IPW, okay? For complete case, we have weight to correctify, okay? And do this weight actually, is coming from two parts. And one part is from the missingness probability. The other part is from the data distribution. Now, the weight actually does not distinguish the missingness probability and the data distribution. The way it treats them equally. And another note I want to say is step two and four are based on the complete case only. Now, let’s look at step four. Okay? Let’s look at step four.
In step four, we saw assumptions are missing at random.

It’s easy to verify this, okay?

Like \(wx\), which is the inverse of the missingness probability

\[ b(X) - E\{b(X)\} | R-1 = 0, \text{ okay?} \]

And in this case, we can let \(b(X)\) to be the score function.

The probability are conditional estimation and the conditional probability under this density.

We can easily write a sample case, a sample scenario.

All the weight is inactive.

Some weight is one,

And this is the estimating equation part,

estimation equation part.

As you can see,

this is a typical empirical likelihood scenario.

So, this is a typical formulation for empirical likelihood.

And the solution actually can be even as in all formula,

our previous, can be given by this one, okay?

The weight can be determined by this.

And though hard, can be estimated by solving this equation.

Okay?

So, that’s all key steps for this methodology, okay?
This actually, is the formula we first written down on the paper. And then we thought, "Okay, this might also be able to be applied to the other scenario." Indeed it can be applied in other scenarios. For example, in this quantile regression with missing covariates. In this scenario, all parameter of interest is 0. This 0 is coming from these linear regression. We want to estimate this 0. And all covariates had two paths, X1 and X2. This X1 path is always observed, while this X2 may have some missing. So, the observed data. And I need copies of this format. This missingness response completely observed covariates and some covariates are missing, some covariates are observed, okay? In this setting, we want to estimate 0, as in previous scenario. We have two sets of models, okay? One set model is for the missing probability. And the other set of model is for data distribution. Here the distribution is related to X2, given the condition of the response and completely of the X1. Now, as previous, we have five steps.
Step one and step two are same as in case one.
And in step one, we estimate in the missing probability.
In step two, we estimate the data distribution.
And then in step three, we get preliminary imputation estimate of 0 by solving this seemed a very complicated equation.
And here there’s Xi, which had two parts, the complete the case and on the missing part.
The missing part is random drawn from this data distribution.
And then the step four, okay?
The key is that the empirical likelihood part where we used to compute to the weight.
And these weights that I had, is for complete case.
And at previous, this weight depends on three parts.
One is missing probability, 1 is the distribution.
Gamma previous, it depend on the preliminary estimate of margin quantile.
Now, it’s related to the preliminary estimate of linear quantile coefficient.
Okay?
After we estimate these weight WI, then we can go to the estimating equation part, okay?
Let’s say five steps. Let’s say five steps.
As you can see, step one, step two, step three,
is all preexisting method we adapt trying to estimate
the missing probability, the data distribution, and also impute to get a preliminary estimate of the parameter we are increasing.

And then from all these, we pull all this information together to get a good weight for the compete case. And then the using this empirical likelihood method

And then we adjust this complete case with the estimated weight to get a final estimate, to get the final multiple robust estimate.

In the case three, the parameter we are interested is still 0.

This linear quantile regression are here. The scenario is the full-data vector is (Y, X). In this scenario, Y is missing and random, okay?

Of course the simple complete a case analysis where lead to a consistent estimate, but it doesn’t mean it will be optimal. Here we are trying to get a more complete educated

but still very practical method.

We are having some auxiliary variable. As this auxiliary variable, usually not the main study interest,
and thus do not enter the quantile regression model. However, we can use it to help us to explain the missingness mechanism and to help us to build a more plausible model for the conditional distribution of Y. Now, here is the observed data. So, we now have an ID copies of these R, RY, this Y gets a missing, X is completely observed, and we have got auxiliary variable S. We have this missing and random scenario. We use \((X, S)\) to denote the probability, and we use \(f(Y \mid X, S)\) to denote conditional density. As previous, we have multiple models for missing probability, and we have multiple models for data distribution. And then once again, we have the all five steps. The first step, we modeled the missing probability. And here we have this additional auxiliary variable. The second step, we model the data distribution. Again, we have this auxiliary variable. And then step three, we get a preliminary estimate on using this imputation method. We have our preliminary estimate of the parameter we are interested in,
which is a linear regression coefficient here.

And then after the preparation of step one, step two, and step three, we finally be able to estimate our weight, okay?

Our weight is for complete case. And from the formula here, you can tell why I put this scenario as scenario three because it got more and more complicated.

Although the weight still depends on three parts, related to the first three step.

The missing probability related to this alpha, the data distribution related to this gamma, and the preliminary estimate made by using imputation in step three.

Our method all framework in general, these five sets, the key thing is step four is empirical likelihood method to estimate the weight.
and we will estimate our framework in these three scenarios.

Of course there are some other scenarios, and you can easily adapt to these five steps. Now, let’s look at some theoretical proprietary.

Why we propose these seemingly complicated five steps.

We first look at the case one. There are two parts.

The first theorem is about this consistence. The second theorem is about asymptotic normality, okay?

So, under certain conditions, if...

Remember we have two sets of models. One set of model, we modeled the probability. The other set of model, we modeled the data distribution.

So if either one from the model of modeling missingness probability is correctly specified, Okay?

Then, our estimate will be consistent. Our estimate it well be consistent.

So, all proposed method allow you to make mistakes, okay?

But you at least make one good right decision, then you get a consistent result, okay?

Of course if you make all the bad decisions, you didn’t choose any track modeling, these two sets of model, then you probably won’t be able to get that consistent result.
And then the second theorem is about the asymptotic normality. Under certain conditions, the model estimate some multiple robust estimate on the marginal quantile where I have asymptotic normal distribution with mean zero and variates here is related to this variable. Variates is related to this data one random variable. And as you can see these variates of data one actually coming from these three parts, the estimate of the missingness probability, the estimate of these data distribution, and also the imputation process, okay? That’s for case one. Similarly for case two, we have these two theorem. Y is consistent. And as long as the one model is correctly specified, we would have this consistency. And then this asymptotic normality, we would have asymptotic normal distribution. And also the variates, they’re two, as you can see. The two is ready to first three step to estimate the different component, okay? And then case three, two theorem. Consistency, we need at least one model. As long as one model is correctly specified, we have a consistent result.
And we have this asymptotic normalcy and the variates come from their three part. Okay?

As you can see, this is a very complicated formula.

It’s a model getting more and more complicated.

And also, if you see that you can compound the variates of the three to the situation with complete case analysis.

Because for complete case analysis, we also get the consistent result, but like I said, it doesn’t mean the variates would be optimal.

And here, we actually can verify the variates of the three will be smaller if our model are correctly specified, okay?

Let’s say theoretical propriety.

Now, let’s look at some simulation, okay?

We did simulation for each scenario, but due to the timely meet, I will only present two.

In the second scenario, we have four here.

We have X1 follow exponential distribution, X2  normal distribution. And so Y is discrete, one is continuous, okay?

The model is the simple linear model and the error distribution Y, as you can see, is heteroscedastic.
Because of these error distribution, it’s reduced to X1.

The missing mechanism for X2, in the second scenario, we have a part of X2 is missing is through this logistic regression, okay?

Now, missingness rate is about 38%.
Eventually, they have this conditional quantile regression,
linear regression, they have those coefficient excess.
This is our simulation setup is in the second scenario.
Now, we consider two working models for
The fist one is correct. The second one is incorrect.
We can see there are two models for the distribution, okay?
All right.
This is the incorrect one
and for the ordinary least squares regression.
And this is correct one with title 0.25 0.75.
We have replication, 1,000 times.
We have some equals 500, L is 10.
This L is really related to the first step
of the imputation.
Okay.
Now, here is all our simulation result, okay?
Although the result has to be multiplied by 100,
as you can see Y is very large.
And also we denote our mass as 0000, okay?
The fist two digit represent
The missing probability model. The last two is data distribution. For example, for IPW 1000, that means we only use inverse probability method. And the weight is estimating is based on this correct weight, okay? And for the imputation, that means we only use this data distribution. And for this IM 0010, that means we use our first model, which is to model the data distribution. This is the second model for data distribution. And in either case, is always the first one is correct model. The first one is correct model. The second one is not, okay? That’s just from notation. As you can see here using IPW if the model is correctly specified, the bias is quite small and everything is quite good. However, if you miss specify the missingness probability, we see the estimate is quite out of control, okay? Let’s say for IM imputation, if you specify correctly the data distribution, the result is good. If not, then it’s not. Okay. Then there’s multiple robust method.
In the multiple robust method, we look at, for example, this one, we get a missing probability correctly specified, then we get a good result. If not, we get bad result as the IPW, okay? But anyway, if we can choose to use all these four models, as you can see, the result is quite good, okay? The taking home method for these simulation study is, if you have some ideas about missingness probability about the state of this data distribution, and you think, "Okay, maybe this one is right or maybe this one is also right, okay? So on my side, just tell you, "Okay, I don’t have to just put all these potential candidate potential model into all framework. Then we look at the recount. This one of the simulation is scenario two. We also have a simulation in a scenario three, but I will skip it here and go directly to the real data analysis. So, in this real data analysis, we look at this AIDS clinical Trials Group Protocol 175 or ACTG 175 data. In this research, we evaluate treatment with either a single nucleosides or through HIV-infected subject whose CD4 cells count and are from 200 to 500 per cubic millimeters.
So, we consider to arms or treatment. One is standardized, and the other one is with three newer treatments.

The two arms respectively, have about 500 and 1,600 subjects. Now, model we are looking at is the linear quantile regression model. The data can be found in this package. Now for the data, the average subject is 35 years old, standard variation is about nine, and the variable CD4 96 is missing for approximate 37%.

It’s quite similar to simulation scenario. Each athlete is part of setup of simulations scenario. However, at baseline during the followup, full measurements on additional variable are correlated with CD4 96 are obtained. So this would be the missing part. We get the missing part. Here we assumed this CD4 96 is the missing and random.

And we also have other baseline, for example, CD4 80 and CD4 20, and so on. We will use these as auxiliary variables. So, we have our third scenario in this real data analysis.

And why we choose this data?
If we look at this CD4 96, the histogram of this, okay?
The left one is before we do it's original skill. The right one is after we do log transformation. So, as you can see, the left one is kind of truncated, and the right one also truncated.

So you may debate, "Okay, which one I should use? Do I take log transformation or not? Or to be, or not to be.”

So that's no apparent reason to favor one of them for the imputation method.

Now, what do we do? In our proposed method, we don't need to make the choice. And because no apparent reason, we take a log, or not take log.

Now, let’s put the two together into our model, okay?

So we can simultaneously accommodate both simulation.

And then we have a eight covariates and auxiliary variable.

Then we have this probability is modeled by logistic regression containing all main effect of X and S.

So, here is the result. Here is the result.

This is a big table, but let me summarize these table.

Okay.
They three newer treatment, significantly slow the progress.

Our proposed method and the IPW method, produce very similar results, okay

And the incubation estimate, one failed to catch difference in the treatment and treatment arm effect for different quantile.

The amputation estimator 2 gives an increasing estimation effect and covariance.

In addition, the two imputation estimates are quite sensitive to the selection of the working models.

Okay?

And also, from these real data, we can help complete case analysis overestimate the treatment arm effects once again,

so that even sometimes the complete case analysis is valid but there are also advantage to use our proposed method.

All right, so here’s the summary of my talk.

We proposed a general framework for quantile estimation with missing data.

And we actually applied these framework in different scenario.

Now, the taking home message is,

our proposed method or whatever robust against possible model misspecification.

So, as we have two sets of model,

one for missing probability

and one is for data distribution.
As long as one model is correct, then we will get good result. And also our method can be easily to be generalized to many other scenario. And I think that’s all of my talk, and thank you.

Thank you, Linglong. This was very interesting. I think we’re almost out of time, so if there’s we have time probably for one question. So if there’s any, if not I believe it will be interesting to see what cases, like what are the conditions for overestimation or underestimation for complete case analysis, I guess.
I guess, it must depend on the data distribution and the missingness mechanism that's been made. But I'm not sure one.

I agree with you. The reason I would answer I don’t know, because it’s really hard to know how the data is miss.

Although we assume it’s missing at runtime. Yeah.

But, who knows the reality?

Right. Yeah, right.

I guess, under your assumption of missing at random, then I guess there could be conditions for underestimation or overestimation under the assumption of where MI.

But, I don’t know.

I was wondering if people have derived those or not.

They could be future work, right?

All right.

Linglong, thank you.

I’ll see you in an hour for a one on one meetings.

And I’ll see you later. All right.
862 01:01:56.558 --> 01:01:57.391 - Thank you.
863 01:01:57.391 --> 01:01:58.224 - Bye-bye. Thank you everyone for joining.
864 01:01:58.224 --> 01:01:59.200 Bye.
865 01:01:59.200 --> 01:02:00.033 - Bye.
866 01:02:00.033 --> 01:02:00.866 - Bye.