WEBVTT

1 00:00:00.030 --> 00:00:01.580 - Hi, everyone.

2 00:00:01.580 --> 00:00:04.400 Welcome to the departmental seminar of

3 00:00:04.400 --> 00:00:07.633 the Departmental Biostatistics, Yale University.

4 00:00:08.800 --> 00:00:12.340 I'm pleased to introduce you Linglong Kong.

5 00:00:12.340 --> 00:00:15.570 He was associate professor of the Department of Mathematical

6 00:00:15.570 --> 00:00:19.880 and Statistical Sciences at the University of Alberta.

7 00:00:19.880 --> 00:00:23.540 He's research interests are on, and correct me if I'm wrong,

8 00:00:23.540 --> 00:00:27.197 on functional and neuro imaging data analysis,

9 00:00:27.197 --> 00:00:28.670 statistical machine learning,

10 00:00:28.670 --> 00:00:32.350 and robost statistics and quantile regression.

11 00:00:32.350 --> 00:00:35.060 So today, he is gonna talk about his work on

12 00:00:35.060 --> 00:00:38.110 general framework for quantile estimation

13 00:00:38.110 --> 00:00:39.423 with incomplete data.

14 00:00:40.400 --> 00:00:43.273 Thank you, Linglong. And whenever you're ready.

15 00:00:44.240 --> 00:00:47.100 - Thank you Laura for the introduction.

16 00:00:47.100 --> 00:00:51.483 And also thanks Professor John for the invitation.

17 00:00:52.320 --> 00:00:56.680 I'm very happy to be here, although it's way too early.

18 00:00:56.680 --> 00:01:00.980 So today I'm going to talk about general framework for

19 00:01:00.980 --> 00:01:04.033 quantile estimation with incomplete data.

20 00:01:13.161 --> 00:01:16.661 So, this is a joint work with Peisong from

21 00:01:20.080 --> 00:01:22.840 University of Michigan and Jiwei from

22 00:01:22.840 --> 00:01:27.130 University of Wisconsin-Madison, and Xingeai.

23 00:01:27.130 --> 00:01:32.130 And we started this work when at the second year

24 00:01:33.180 --> 00:01:36.353 when I started my position at the University of Alberta.

25 00:01:37.370 --> 00:01:42.370 I know Peisong a long time ago before he was a student,

26 00:01:43.730 --> 00:01:47.600 and at that time he just started his position as

27 00:01:47.600 --> 00:01:51.050 assistant professor at the University of Waterloo.

28 00:01:51.050 --> 00:01:56.050 And I invited him to visit me and afterwards,

29 00:01:56.248 --> 00:01:58.400 he invited me to visit him.

30 00:01:58.400 --> 00:02:02.040 And we feel like we visited each other already,

31 00:02:02.040 --> 00:02:04.140 we should get something done.

32 00:02:04.140 --> 00:02:09.140 But I remember that I've known where he stayed in his office

33 00:02:10.910 --> 00:02:14.500 at the University of Waterloo and thinking about

34 00:02:14.500 --> 00:02:17.070 what do we have to do together.

35 00:02:17.070 --> 00:02:19.570 And eventually we thought, ”Okay, what I'm good at

36 00:02:20.550 --> 00:02:23.780 and while all my research area is quantile regression.

37 00:02:23.780 --> 00:02:25.693 And what is Peisong good at?

38 00:02:26.675 --> 00:02:31.410 One of the research area of Peisong is missing the data.”

39 00:02:31.410 --> 00:02:34.220 So we said maybe we can put them together,

40 00:02:34.220 --> 00:02:39.220 then we are write a couple of formula on the paper.

41 00:02:40.870 --> 00:02:44.590 Then we feel like, ”Okay, we get a copy already.”

42 00:02:44.590 --> 00:02:47.630 Then we went to have a dinner.

43 00:02:47.630 --> 00:02:52.473 And then one year later Peisong send me like

44 00:02:52.473 --> 00:02:57.330 two pages to trap, said maybe we should continue it.

45 00:02:57.330 --> 00:03:02.330 And that's the first scenario in this topic,

46 00:03:02.620 --> 00:03:04.200 I'm gonna talk about.

47 00:03:04.200 --> 00:03:09.200 And then another half year, I sent him my feedback.

48 00:03:11.870 --> 00:03:15.050 I said, ”Why don't we make it more general,

49 00:03:15.050 --> 00:03:16.880 make it a framework?”

50 00:03:16.880 --> 00:03:19.940 So this semester we’re going to be able to apply

51 00:03:19.940 --> 00:03:22.350 to honor other scenarios.

52 00:03:22.350 --> 00:03:25.980 And then we we both feel it’s good idea,

53 00:03:25.980 --> 00:03:27.080 then we started working on it.

54 00:03:27.080 --> 00:03:31.360 At that time, Jiwei was posed to at a University of Waterloo

55 00:03:32.760 --> 00:03:34.980 and Xingeai where my post are.

56 00:03:34.980 --> 00:03:38.160 So, we thought together and started a project.

57 00:03:38.160 --> 00:03:43.160 Eventually, I wound a project that I’m kind of proud of.

58 00:03:46.840 --> 00:03:49.220 So, what’s missing data?

59 00:03:49.220 --> 00:03:51.900 The missing data arise in almost all

60 00:03:51.900 --> 00:03:53.703 serious statistical analysis.

61 00:03:55.600 --> 00:03:59.287 Missing on values are representative of the

62 00:04:02.633 --> 00:04:03.983 messiness of real world.

63 00:04:04.950 --> 00:04:07.700 Why we would have missing a missing value,

64 00:04:07.700 --> 00:04:10.793 it could be all kinds of reason.

65 00:04:11.710 --> 00:04:16.610 For example, it may be due to social or natural process.

66 00:04:16.610 --> 00:04:20.330 Like for example, a student get a graduate,

67 00:04:20.330 --> 00:04:25.330 get a job out in clinical trial, people get died, and so on.

68 00:04:26.290 --> 00:04:28.720 And also could happen that you survey.

69 00:04:28.720 --> 00:04:31.600 For example, in certain question asked,

70 00:04:31.600 --> 00:04:34.720 only asked respondent answer yes,

71 00:04:34.720 --> 00:04:37.003 to continue to answer certain questions.

72 00:04:38.090 --> 00:04:41.360 Or maybe it’s the intention missing

73 00:04:41.360 --> 00:04:43.353 as a part of a data collection process.

74 00:04:44.580 --> 00:04:48.100 Or some other scenario including random data collection

75 00:04:48.100 --> 00:04:52.483 issues respondent refusal or non-response.

76 00:04:56.120 --> 00:05:01.120 So, mathematically how we categorize these kind of missing,

77 00:05:01.150 --> 00:05:05.020 and here is the three scenario.

78 00:05:05.020 --> 00:05:08.513 Now, first scenario we call it missing completely at random.

79 00:05:09.740 --> 00:05:10.730 What does that mean?

80 00:05:10.730 --> 00:05:14.790 That means the missingness is nothing to do with the

81 00:05:14.790 --> 00:05:15.840 person being studied.

82 00:05:16.920 --> 00:05:19.024 They're just completely got missing,

83 00:05:19.024 --> 00:05:21.633 it's nothing related to any feature of this person.

84 00:05:22.840 --> 00:05:25.983 The second scenario is missing at random.

85 00:05:25.983 --> 00:05:30.200 Missing is to do with the person, but can be predicted

86 00:05:30.200 --> 00:05:32.890 from other information about the person.

87 00:05:34.060 --> 00:05:37.733 Like either a certain scenario need these project,

88 00:05:38.641 --> 00:05:43.093 the missingness maybe predictive from some

89 00:05:43.093 --> 00:05:46.013 auxiliary verbals auxiliary information.

90 00:05:48.240 --> 00:05:51.443 The third one is a very hard one, is missing not at random.

91 00:05:55.250 --> 00:05:59.110 The missingness depends on observed the information

92 00:05:59.110 --> 00:06:03.653 and sometime even the response itself.

93 00:06:04.770 --> 00:06:08.390 So, the missingness is specifically related to

94 00:06:08.390 --> 00:06:09.360 what is missing.

95 00:06:09.360 --> 00:06:12.750 For example, a person to not attend a drug test

96 00:06:12.750 --> 00:06:15.403 because the person took drugs the night before.

97 00:06:16.690 --> 00:06:18.280 And therefore the second day,

98 00:06:18.280 --> 00:06:20.380 he couldn't make to the drug test.

99 00:06:20.380 --> 00:06:22.313 Couldn't get to that drug test result.

100 00:06:23.347 --> 00:06:26.363 These are three missing mechanism.

101 00:06:30.360 --> 00:06:33.410 How do we handle those missing data?

102 00:06:33.410 --> 00:06:34.970 There are many strategies.

103 00:06:34.970 --> 00:06:37.300 For example, the first one would be,

104 00:06:37.300 --> 00:06:40.240 well, let's try to get the meeting data.

105 00:06:40.240 --> 00:06:41.540 That would be great.

106 00:06:41.540 --> 00:06:45.480 But in reality, that's usually impossible.

107 00:06:47.560 --> 00:06:51.900 But the second is, well, as we have incomplete cases,

108 00:06:51.900 --> 00:06:54.813 let's just discard.

109 00:06:57.018 --> 00:07:02.018 Just analyze the complete case, right?

110 00:07:02.090 --> 00:07:05.200 But these could cause some other problems.

111 00:07:05.200 --> 00:07:06.313 We will talk about it.

112 00:07:07.180 --> 00:07:11.620 And the third one is we replace missing data

113 00:07:11.620 --> 00:07:14.400 by some conservative estimation.

114 00:07:14.400 --> 00:07:18.463 For example, using sample mean, sample median, and so on.

115 00:07:20.200 --> 00:07:25.150 The first one is we are trying to estimate the missing data

116 00:07:25.150 --> 00:07:26.900 from other data on the person.

117 00:07:26.900 --> 00:07:31.170 We use on sort of more sophisticated method to impute.

118 00:07:37.260 --> 00:07:41.273 Now in particular, mathematically speaking,

119 00:07:43.072 --> 00:07:45.687 the strategy we are using today do to deal

120 00:07:45.687 --> 00:07:47.870 with missing data,

121 00:07:47.870 --> 00:07:50.570 the first one is a complete case analysis.

122 00:07:50.570 --> 00:07:52.310 These are very simple, okay?

123 00:07:52.310 --> 00:07:55.503 We just analyze compete case, okay?

124 00:07:56.360 --> 00:08:00.650 And we only analyze in consideration that individuals with

125 00:08:00.650 --> 00:08:01.713 no missing data.

126 00:08:04.950 --> 00:08:07.150 Sometimes it can provide good result,

127 00:08:07.150 --> 00:08:12.030 but the estimation obtained from this complete case analysis

128 00:08:12.030 --> 00:08:17.030 maybe biased if they excluded individuals are systematically

129 00:08:17.520 --> 00:08:20.290 different from those included.

130 00:08:20.290 --> 00:08:24.410 So hence, if the complete case would be a good

131 00:08:24.410 --> 00:08:28.450 representation of those missing case,

132 00:08:28.450 --> 00:08:33.450 then this method would it be fine.

133 00:08:33.860 --> 00:08:37.800 Otherwise, if the complete case is quite different from

134 00:08:37.800 --> 00:08:42.313 those we miss, then all result can be biased.

135 00:08:44.300 --> 00:08:48.653 And then there's inverse probability weighting method IPW.

136 00:08:49.780 --> 00:08:53.470 This is a commonly use method to correct the bias from a

137 00:08:53.470 --> 00:08:55.063 complete case analysis.

138 00:08:55.900 --> 00:08:56.733 What does that mean?

139 00:08:56.733 --> 00:09:01.660 It means, okay, we give each complete case a weight.

140 00:09:03.230 --> 00:09:07.292 This weight is the inverse of the probability of

141 00:09:07.292 --> 00:09:12.150 being a complete case.

142 00:09:12.150 --> 00:09:14.330 Well, this can also cause some bias

143 00:09:15.810 --> 00:09:19.833 if this IPW relies on the data distribution.

144 00:09:25.490 --> 00:09:28.940 The first strategy is more sophisticated to do

145 00:09:28.940 --> 00:09:31.000 these multiple imputation.

146 00:09:31.000 --> 00:09:32.260 It's quite common method,

147 00:09:32.260 --> 00:09:35.192 especially nowadays in genetic study.

148 00:09:35.192 --> 00:09:39.360 How do we do multiple imputation?

149 00:09:39.360 --> 00:09:43.730 We create multiple sets of imputation for

150 00:09:43.730 --> 00:09:48.070 the missing values, using imputation process

151 00:09:48.070 --> 00:09:49.693 with a random component.

152 00:09:50.900 --> 00:09:53.560 Now, we have an full data set.

153 00:09:53.560 --> 00:09:58.560 Then we analyze each data set.

154 00:09:58.860 --> 00:10:02.300 Those full data set can be a little bit different.

155 00:10:02.300 --> 00:10:07.300 Can be slightly different because the randomness of

156 00:10:07.900 --> 00:10:09.773 the imputation process.

157 00:10:10.720 --> 00:10:13.540 Anyway, analyze those data set, complete the data set,

158 00:10:13.540 --> 00:10:17.023 and then we get all set of parameter estimates.

159 00:10:17.023 --> 00:10:19.770 Then we can combine those result.

160 00:10:19.770 --> 00:10:21.273 We can combine this result,

161 00:10:22.361 --> 00:10:24.473 and we hopefully we get a better result.

162 00:10:26.065 --> 00:10:29.823 The multiple imputation sometimes works quite well,

163 00:10:31.030 --> 00:10:35.000 but only if the missing data can be ignored.

164 00:10:35.959 --> 00:10:39.304 And also, we have a good imputation models.

165 00:10:39.304 --> 00:10:41.290 And while it depends on the nature of the data,

166 00:10:41.290 --> 00:10:44.551 the auto mind depends on what kind of imputation model

167 00:10:44.551 --> 00:10:46.023 we are going to use.

168 00:10:51.380 --> 00:10:54.853 Now, that's how we deal with missing data,

169 00:10:56.040 --> 00:11:00.033 the strategy we happen to use to deal with missing data.

170 00:11:01.000 --> 00:11:06.000 But let's matched them together in terms of missing data.

171 00:11:06.460 --> 00:11:10.720 How we use these meeting dates age to deal with

172 00:11:10.720 --> 00:11:12.703 different missing mechanism.

173 00:11:13.660 --> 00:11:17.773 For example, if the data is missing complete at random,

174 00:11:18.720 --> 00:11:23.293 now in this case, the complete case analysis is quite good.

175 00:11:25.230 --> 00:11:29.200 Multiple imputation or any other imputation methods

176 00:11:29.200 --> 00:11:30.520 is also okay.

177 00:11:30.520 --> 00:11:31.750 Is also valid.

178 00:11:31.750 --> 00:11:35.530 So, this missing complete at random is

179 00:11:35.530 --> 00:11:38.290 the easiest case to deal with.

180 00:11:39.890 --> 00:11:42.930 What if data is missing at random?

181 00:11:42.930 --> 00:11:47.930 Then in this case, some complete case analysis are valid

182 00:11:51.250 --> 00:11:55.740 and multiple imputation nearly is okay too,

183 00:11:55.740 --> 00:11:57.993 if the bias is negligible.

184 00:11:59.720 --> 00:12:02.080 Now in a certain case,

185 00:12:02.080 --> 00:12:05.300 if the data is missing not at random,

186 00:12:05.300 --> 00:12:09.643 then we have to model the missingness explicitly.

187 00:12:11.230 --> 00:12:14.520 We need jointly modeling the response.

188 00:12:14.520 --> 00:12:16.780 We need jointly model the response,

189 00:12:16.780 --> 00:12:19.313 and also the missingness.

190 00:12:21.769 --> 00:12:23.079 In practice of course,

191 00:12:23.079 --> 00:12:28.079 we try to assume missing and random whenever it’s possible

192 00:12:28.160 --> 00:12:31.560 and try to avoid to deal with

193 00:12:31.560 --> 00:12:34.010 missing not at a random situation.

194 00:12:34.010 --> 00:12:39.010 But the reality, it’s not anything that we can control.

195 00:12:40.720 --> 00:12:45.240 Sometime we have data always missing not either random.

196 00:12:45.240 --> 00:12:50.240 Think in that case center or there is one special issue

197 00:12:52.960 --> 00:12:56.623 dedicated to missing data, not at a random situation.

198 00:13:01.750 --> 00:13:03.450 Now, we have different strategies.

199 00:13:04.380 --> 00:13:06.670 And that they state different strategies

200 00:13:06.670 --> 00:13:11.670 have different advantage and disadvantage.

201 00:13:12.370 --> 00:13:17.188 For example, multiple imputation is generally more efficient

202 00:13:17.188 --> 00:13:21.393 than IPW, but it's more complex.

203 00:13:22.880 --> 00:13:26.760 And the imputation and IPW approach

204 00:13:28.239 --> 00:13:32.433 require to model the data distribution

205 00:13:32.433 --> 00:13:34.930 and the missingness probability, respectively.

206 00:13:34.930 --> 00:13:38.550 Imputation, we need to model data distribution.

207 00:13:38.550 --> 00:13:43.183 IPW, we need model the missingness probability.

208 00:13:45.154 --> 00:13:48.164 And also, for all kinds of strategy,

209 00:13:48.164 --> 00:13:51.810 we would have have good property,

210 00:13:51.810 --> 00:13:56.163 only if the corresponding model is correctly specified.

211 00:13:59.030 --> 00:14:03.220 Most existing method are vulnerable to

212 00:14:03.220 --> 00:14:06.098 these model misspecifications.

213 00:14:06.098 --> 00:14:10.670 Of course can use nonparametric method to reduce the risk

214 00:14:10.670 --> 00:14:15.670 of model misspecification, but it's often impractical

215 00:14:16.040 --> 00:14:18.523 due to the curse of dimensionality.

216 00:14:21.200 --> 00:14:26.200 So now, how do we deal with this model misspecification?

217 00:14:27.012 --> 00:14:30.370 We have some method available.

218 00:14:30.370 --> 00:14:35.313 For example, we can use a double robust method.

219 00:14:36.900 --> 00:14:39.900 In particular, in double robust method,

220 00:14:39.900 --> 00:14:41.913 we have this augmented IPW.

221 00:14:44.300 --> 00:14:49.200 We are not only model the missingness probability,

222 00:14:49.200 --> 00:14:51.137 but also the distribution.

223 00:14:52.210 --> 00:14:54.410 Why is double robust?

224 00:14:54.410 --> 00:14:57.930 Because the result would be confusing

225 00:14:57.930 --> 00:15:00.110 if the model is correct.

226 00:15:02.160 --> 00:15:05.860 If the way we model missingness probability

227 00:15:06.774 --> 00:15:11.540 or the way we model the distribution is correct,

228 00:15:11.540 --> 00:15:14.467 then we would get consistent result.

229 00:15:14.467 --> 00:15:16.517 And that's why it's called double robust.

230 00:15:17.910 --> 00:15:21.530 Well, now that we are not satisfied with double robust,

231 00:15:21.530 --> 00:15:25.290 what about we can a multiple guarantee?

232 00:15:25.290 --> 00:15:27.203 So, we have these multiple robust.

233 00:15:27.203 --> 00:15:30.883 This is a proposal by Peisong.

234 00:15:32.560 --> 00:15:37.560 And they multiple robust method is proposed to account for

235 00:15:37.990 --> 00:15:42.100 multiple models for missingness probability

236 00:15:42.100 --> 00:15:43.413 and the distribution.

237 00:15:45.024 --> 00:15:48.296 In double robust, we can only one model for missingness

238 00:15:48.296 --> 00:15:51.370 probability and one model for data distribution.

239 00:15:51.370 --> 00:15:52.670 Well, for multiple robust,

240 00:15:53.580 --> 00:15:57.563 we get multiple models to model missingness probability,

241 00:15:58.810 --> 00:16:03.027 and we can have multiple models to model distribution.

242 00:16:04.670 --> 00:16:09.670 The good thing is the estimation result will be consistent

243 00:16:10.822 --> 00:16:15.713 if either one or the model is correct.

244 00:16:18.970 --> 00:16:23.243 Now, let's look at those crushing mathematically.

245 00:16:25.780 --> 00:16:29.340 So, we are looking at missing at random.

246 00:16:29.340 --> 00:16:33.520 We assume on the observed data are ID.

247 00:16:33.520 --> 00:16:36.217 So we have data R, RY XT.

248 00:16:37.673 --> 00:16:41.940 R, we use it to missingness, and the IPW estimator,

249 00:16:47.730 --> 00:16:52.470 essentially we are trying to solve these equation.

250 00:16:52.470 --> 00:16:56.323 And here, these is the probability,

251 00:16:57.770 --> 00:17:01.200 although makes complete case.

252 00:17:01.200 --> 00:17:02.980 And IPW is consistent,

253 00:17:02.980 --> 00:17:06.503 only if this X is correctly specified.

254 00:17:08.330 --> 00:17:10.490 And then, then from the equation,

255 00:17:10.490 --> 00:17:13.132 we can get consistent estimate of those

256 00:17:13.132 --> 00:17:15.465 permit we are interested in.

257 00:17:17.057 --> 00:17:20.474 This is IPW. The other one is imputation.

258 00:17:23.377 --> 00:17:27.510 For imputation, we need model that take distribution.

259 00:17:27.510 --> 00:17:32.510 And here we have on the model of a f(Y|X)

260 00:17:35.853 --> 00:17:36.870 And as you can see,

261 00:17:36.870 --> 00:17:41.870 we have our imputation for those missing data.

262 00:17:43.730 --> 00:17:47.003 This imputation is consistent,

263 00:17:47.003 --> 00:17:51.890 only if this state distribution is correctly modeled,

264 00:17:51.890 --> 00:17:55.293 this f(Y|X) is correctly modeled.

265 00:17:58.240 --> 00:18:03.240 Now for these augmented inverse probability waited method,

266 00:18:04.950 --> 00:18:09.950 we actually combined these two together.

267 00:18:10.950 --> 00:18:13.900 We had the first part from IPW,

268 00:18:13.900 --> 00:18:16.610 second part from implication.

269 00:18:16.610 --> 00:18:21.610 So the estimation result would be consistent

270 00:18:22.640 --> 00:18:27.640 if either this model for missingness probability

271 00:18:28.030 --> 00:18:32.633 or the model for data distribution is correctly specified.

272 00:18:34.820 --> 00:18:38.209 Well, for multiple robust method,

273 00:18:38.209 --> 00:18:43.209 they have a serious model for missingness probability

274 00:18:43.670 --> 00:18:47.163 and a serious model for data distribution.

275 00:18:48.790 --> 00:18:53.070 And all result would be consistent,

276 00:18:53.070 --> 00:18:55.843 if any one model is correctly specified.

277 00:19:00.930 --> 00:19:02.860 Well, this is something

278 00:19:02.860 --> 00:19:06.033 I just get a quick review about this missing data.

279 00:19:06.900 --> 00:19:09.760 Like I said, this is the part Peisong is

280 00:19:11.570 --> 00:19:13.680 one of the Peisong research area.

281 00:19:13.680 --> 00:19:18.290 For me, my research area is quantile regression.

282 00:19:18.290 --> 00:19:23.030 So, internal quantile regression at that time

283 00:19:23.030 --> 00:19:25.750 we were thinking, ”Okay, those methods,

284 00:19:25.750 --> 00:19:30.750 these IPW, AIPW or double robust method,

285 00:19:31.590 --> 00:19:35.120 multiple robust method, had been quite well studied

286 00:19:35.120 --> 00:19:39.108 for when we model the conditional mean.

287 00:19:39.108 --> 00:19:41.160 Therefore, condition of quantile, there are not

288 00:19:41.160 --> 00:19:42.833 a lot of methods available.

289 00:19:44.320 --> 00:19:46.307 Why we care about the quantile?

290 00:19:46.307 --> 00:19:48.720 A quantile not only provide a central feature

291 00:19:48.720 --> 00:19:53.043 of the distribution, but also care about the tail behavior.

292 00:19:57.290 --> 00:20:00.690 And also under very mild conditions,

293 00:20:00.690 --> 00:20:04.510 the quantile function can uniquely determine

294 00:20:04.510 --> 00:20:05.910 the underlying distribution.

295 00:20:07.440 --> 00:20:12.440 So, there are a lot of advantages to model the quantiles.

296 00:20:12.550 --> 00:20:17.550 Then, we decided to study these missingness

297 00:20:17.640 --> 00:20:19.493 in quantile estimation.

298 00:20:20.550 --> 00:20:23.160 In particular, we proposed a general framework

299 00:20:23.160 --> 00:20:26.273 for quantile estimation with missing data.

300 00:20:29.940 --> 00:20:34.740 So, our proposed model, these kind of framework,

301 00:20:34.740 --> 00:20:38.200 can do a lot of estimation for

302 00:20:38.200 --> 00:20:41.083 missingness in quantile estimation.

303 00:20:42.820 --> 00:20:45.570 But in this paper,

304 00:20:45.570 --> 00:20:50.153 we particularly applied all proposed method,

305 00:20:50.153 --> 00:20:51.203 these three scenario.

306 00:20:52.410 --> 00:20:56.370 Okay, three commonly encountered situation.

307 00:20:56.370 --> 00:21:01.000 The first one we trying to estimate

308 00:21:01.000 --> 00:21:03.193 the marginal quantile of response.

309 00:21:04.280 --> 00:21:08.570 This response get some missingness.

310 00:21:08.570 --> 00:21:11.473 Well, there are fully observed covariates.

311 00:21:12.720 --> 00:21:16.150 That's the first scenario, response gets some missingness

312 00:21:16.150 --> 00:21:20.310 while the corresponding covariates get fully observed.

313 00:21:20.310 --> 00:21:22.810 The second scenario, we are looking at

314 00:21:22.810 --> 00:21:26.803 the conditional quantile of a fully observed response.

315 00:21:27.963 --> 00:21:30.690 In this scenario, we look at

316 00:21:30.690 --> 00:21:35.540 there are some covariates are partialy available.

317 00:21:35.540 --> 00:21:37.313 So, we have some missingness for covariates.

318 00:21:38.900 --> 00:21:42.950 And then the third scenario, we are still looking at

319 00:21:42.950 --> 00:21:45.933 the conditional quantile of a response.

320 00:21:47.380 --> 00:21:52.360 And in this case, the response gets some missingness

321 00:21:52.360 --> 00:21:55.290 and we have fully observed covariates

322 00:21:55.290 --> 00:21:58.393 and also extra auxiliary variable.

323 00:22:02.450 --> 00:22:07.145 Now, let's look at the first situation.

324 00:22:07.145 --> 00:22:09.883 We want to estimate the marginal quantile.

325 00:22:09.883 --> 00:22:14.883 In this scenario, we have the response gets some missingness

326 00:22:17.900 --> 00:22:20.233 and we have the covariates fully observed.

327 00:22:22.050 --> 00:22:25.820 Now, let m to be the number of subjects with

328 00:22:25.820 --> 00:22:29.143 data completely observed.

329 00:22:29.980 --> 00:22:34.980 Then our method consists of the following five steps.

330 00:22:38.104 --> 00:22:42.950 The first step, we calculate this or estimate to this .

331 00:22:45.443 --> 00:22:49.453 This isn't related to the missingness probability, okay?

332 00:22:51.920 --> 00:22:56.920 The way we estimate this, is by maximizing

333 00:22:57.440 --> 00:22:59.193 the binomial likelihood.

334 00:23:00.570 --> 00:23:03.410 So, the first step we estimate the ,

335 00:23:03.410 --> 00:23:08.410 and then we get estimate of the missingness probability.

336 00:23:09.680 --> 00:23:10.600 Okay?

337 00:23:10.600 --> 00:23:13.783 The second step, we calculate gamma.

338 00:23:16.053 --> 00:23:20.740 This gamma is related to this data distribution.

339 00:23:20.740 --> 00:23:25.200 So, we maximize this data distribution.

340 00:23:25.200 --> 00:23:29.310 This gamma is a parameter related to the distribution.

341 00:23:33.254 --> 00:23:36.004 And then the third step is we can

342 00:23:39.352 --> 00:23:43.231 sort of preliminary estimate of the quantile

343 00:23:43.231 --> 00:23:48.064 or the marginal quantile through these imputation process,

344 00:23:51.150 --> 00:23:53.293 by solving this equation.

345 00:23:54.960 --> 00:23:59.733 And as you can see this is quite close to the AIPW scenario.

346 00:24:04.880 --> 00:24:05.713 Okay?

347 00:24:05.713 --> 00:24:10.620 And in this equation, this five is the score function

348 00:24:12.610 --> 00:24:15.883 of quantile lost function.

349 00:24:17.170 --> 00:24:21.647 This prosaic is r - i(r<0).

350 00:24:23.018 --> 00:24:27.930 This is the generalized derivative

351 00:24:27.930 --> 00:24:30.773 of quantile lost function, okay?

352 00:24:33.880 --> 00:24:38.880 Here, this one can not be exact zero.

353 00:24:39.290 --> 00:24:44.290 The reason this phosaica is a non-smooth function.

354 00:24:46.160 --> 00:24:51.160 and it sometime it won't be exact here.

355 00:24:53.100 --> 00:24:55.773 Basically the first step, okay?

356 00:24:57.060 --> 00:25:00.790 Now, we have a preliminary estimator

357 00:25:00.790 --> 00:25:02.910 of the marginal quantile.

358 00:25:02.910 --> 00:25:07.910 The first step is the case that of method

359 00:25:08.060 --> 00:25:11.713 is where the multiple robustness is coming from.

360 00:25:14.070 --> 00:25:18.650 Now, we calculates weights for the complete case.

361 00:25:18.650 --> 00:25:20.860 In total, do we have m complete case.

362 00:25:20.860 --> 00:25:23.640 For each case, we calculate the weight.

363 00:25:23.640 --> 00:25:28.640 As you can see, the weight is determined by three parts.

364 00:25:32.320 --> 00:25:35.790 The first part is related to this alpha,

365 00:25:35.790 --> 00:25:39.023 which is related to the missing probability, okay?

366 00:25:40.330 --> 00:25:41.330 Missing probability.

367 00:25:42.900 --> 00:25:46.063 The second part is related to this gamma.

368 00:25:47.130 --> 00:25:50.103 This is related to the data distribution.

369 00:25:51.590 --> 00:25:56.470 The third part is related to this cube.

370 00:25:56.470 --> 00:26:01.470 This preliminary estimate of these marginal quantile,

371 00:26:01.920 --> 00:26:06.600 which is related to this self step.

372 00:26:06.600 --> 00:26:10.140 As you can see from the first three step,

373 00:26:10.140 --> 00:26:13.730 we are trying to get ready for this,

374 00:26:13.730 --> 00:26:18.434 to get the estimate for the weight for the complete case,

375 00:26:18.434 --> 00:26:19.584 for this complete case.

376 00:26:20.620 --> 00:26:23.300 And also, we have our parameter,

377 00:26:23.300 --> 00:26:27.090 though is obtained through

378 00:26:27.090 --> 00:26:30.713 minimizing these equation, through minimizing this equation.

379 00:26:33.120 --> 00:26:35.570 Now, after we calculate the weight

380 00:26:35.570 --> 00:26:40.570 we get off final estimate of our multiple robust estimate

381 00:26:41.660 --> 00:26:46.487 by solving the following with estimated equation.

382 00:26:49.910 --> 00:26:51.943 This wi is the width.

383 00:26:51.943 --> 00:26:55.360 We estimate it from the first four steps.

384 00:26:57.670 --> 00:27:02.670 And this posy is a score function of quantile loss, okay?

385 00:27:06.240 --> 00:27:10.143 Now, you may get wondering on what's going on

386 00:27:10.143 --> 00:27:13.870 with these five steps.

387 00:27:13.870 --> 00:27:18.870 And let me try to explain it one by one, okay?

388 00:27:19.620 --> 00:27:24.220 In the first step, we get the estimate of alpha, okay?

389 00:27:24.220 --> 00:27:26.503 We get the estimate of alpha.

390 00:27:27.750 --> 00:27:32.750 In sense trying to model they missingness probability, okay?

391 00:27:33.443 --> 00:27:35.259 Missingness probability.

392 00:27:35.259 --> 00:27:40.259 And of course, this missingness probability is consistent

393 00:27:40.703 --> 00:27:45.130 only if this model is correctly specified, okay?

394 00:27:45.130 --> 00:27:48.850 So in the first step, we actually have multiple models

395 00:27:48.850 --> 00:27:52.278 to model the missingness probability.

396 00:27:52.278 --> 00:27:57.278 And you need a hope at least a one model is correct.

397 00:27:57.330 --> 00:27:59.739 Now, in the other case, the missingness probability

398 00:27:59.739 --> 00:28:03.253 will not be correctly specified.

399 00:28:04.610 --> 00:28:08.550 Well, in the second step, we only estimate gamma.

400 00:28:08.550 --> 00:28:10.810 We are trying to model the data distribution

401 00:28:13.585 --> 00:28:17.547 and we have models for the data distribution.

402 00:28:19.860 --> 00:28:21.060 And then the third step,

403 00:28:22.000 --> 00:28:25.570 we are sort of doing some imputation as made

404 00:28:25.570 --> 00:28:28.043 of these marginal quantile.

405 00:28:32.467 --> 00:28:37.467 And these marginal quantile will be correctly estimated,

406 00:28:41.620 --> 00:28:46.123 if those data distribution is correctly specified.

407 00:28:50.240 --> 00:28:52.625 Now for the key staff,

408 00:28:52.625 --> 00:28:54.280 (coughs)

409 00:28:54.280 --> 00:28:55.113 Excuse me.

410 00:28:55.113 --> 00:28:59.370 The step four is typical formulation of

411 00:28:59.370 --> 00:29:02.780 an empirical likelihood program.

412 00:29:02.780 --> 00:29:07.780 I will getting back to this in the next slide,

413 00:29:08.420 --> 00:29:11.840 why it's a empirical likelihood program.

414 00:29:11.840 --> 00:29:16.193 And this is a key contribution of methodology.

415 00:29:17.700 --> 00:29:22.160 Now, in step five, we have the structure of IPW, okay?

416 00:29:23.027 --> 00:29:28.027 For complete case, we have weight to correctify, okay?

417 00:29:31.610 --> 00:29:35.460 And do this weight actually, is coming from two parts.

418 00:29:35.460 --> 00:29:40.460 And one part is from the missingness probability.

419 00:29:40.930 --> 00:29:44.541 The other part is from the data distribution.

420 00:29:44.541 --> 00:29:48.100 Now, the weight actually does not distinguish

421 00:29:48.100 --> 00:29:52.333 the missingness probability and the data distribution.

422 00:29:53.610 --> 00:29:55.253 The way it treats them equally.

423 00:29:59.030 --> 00:30:03.488 And another note I want to say is step two and four

424 00:30:03.488 --> 00:30:07.650 are based on the complete case only.

425 00:30:11.550 --> 00:30:14.515 Now, let's look at step four.

426 00:30:14.515 --> 00:30:17.614 Okay? Let's look at step four.

427 00:30:17.614 --> 00:30:21.393 In step four, we saw assumption are missing at random.

428 00:30:25.890 --> 00:30:28.543 It’s easy to verify this, okay?

429 00:30:28.543 --> 00:30:32.820 Like wx, which is the inverse of the missingness probability

430 00:30:34.300 --> 00:30:39.300 times $b(X) - E\{b(X)\}| R\text{-}1 = 0$, okay?

431 00:30:43.256 --> 00:30:48.200 And in thus case, we can let $b(X)$ to be the score function

432 00:30:48.200 --> 00:30:50.233 of quantile lost function.

433 00:30:51.850 --> 00:30:55.513 And these probability are conditional estimation

434 00:30:55.513 --> 00:30:59.270 and the conditional probability under this density.

435 00:31:00.740 --> 00:31:05.003 And because of this, okay?

436 00:31:06.180 --> 00:31:11.180 We can easily write a sample case, a sample scenario.

437 00:31:13.520 --> 00:31:16.130 So, the scenario is like this.

438 00:31:16.130 --> 00:31:19.230 All the weight is inactive.

439 00:31:19.230 --> 00:31:20.627 Some weight is one,

440 00:31:21.650 --> 00:31:25.351 and this is the estimating equation part,

441 00:31:25.351 --> 00:31:27.434 estimation equation part.

442 00:31:28.536 --> 00:31:30.070 As you can see,

443 00:31:30.070 --> 00:31:35.070 this is a typical empirical likelihood scenario.

444 00:31:40.130 --> 00:31:44.363 So, this is a typical formulation for empirical likelihood.

445 00:31:46.907 --> 00:31:51.423 And the solution actually can be even as in all formula,

446 00:31:55.420 --> 00:32:00.420 our previous, can be given by this one, okay?

447 00:32:01.660 --> 00:32:03.863 The weight can be determined by this.

448 00:32:04.890 --> 00:32:09.890 And though hard, can be estimated by solving this equation.

449 00:32:16.280 --> 00:32:17.113 Okay?

450 00:32:18.840 --> 00:32:23.840 So, that’s all key steps for this methodology, okay?

451 00:32:28.690 --> 00:32:33.690 This actually, is the formula we first written down

452 00:32:34.680 --> 00:32:35.620 on the paper.

453 00:32:35.620 --> 00:32:40.110 And then we thought, "Okay, this might also be able

454 00:32:40.110 --> 00:32:42.767 to be applied to the other scenario."

455 00:32:43.637 --> 00:32:47.840 Indeed it can be applied in other scenarios.

456 00:32:47.840 --> 00:32:52.300 For example, in this quantile regression

457 00:32:52.300 --> 00:32:53.713 with missing covariates.

458 00:32:55.450 --> 00:32:59.713 In this scenario, all parameter of interest is 0.

459 00:33:00.571 --> 00:33:05.250 This 0 is coming from these linear regression.

460 00:33:05.250 --> 00:33:07.213 We want to estimate this 0.

461 00:33:09.726 --> 00:33:14.726 And all covariates had two paths, X1 and X2.

462 00:33:17.120 --> 00:33:19.983 This X1 path is always observed,

463 00:33:21.670 --> 00:33:24.267 while this X2 may have some missing.

464 00:33:26.616 --> 00:33:28.449 So, the observed data.

465 00:33:30.508 --> 00:33:33.340 And I need copies of this format.

466 00:33:33.340 --> 00:33:38.340 This missingness response completely observed covariates

467 00:33:42.510 --> 00:33:44.350 and some covariates are missing,

468 00:33:45.463 --> 00:33:49.100 some covariates are observed, okay?

469 00:33:49.100 --> 00:33:53.173 So, in this setting, we want to estimate 0,

470 00:33:55.020 --> 00:33:59.180 as in previous scenario.

471 00:33:59.180 --> 00:34:02.490 We have two sets of models, okay?

472 00:34:02.490 --> 00:34:07.490 One set model is for , the missing probability.

473 00:34:08.147 --> 00:34:12.633 And the other set of model is for data distribution.

474 00:34:14.910 --> 00:34:19.360 Here the distribution is related to X2,

475 00:34:19.360 --> 00:34:21.440 given the condition of the response

476 00:34:21.440 --> 00:34:23.867 and completely of the X1.

477 00:34:26.860 --> 00:34:31.860 Now, as previous, we have five steps.

478 00:34:34.818 --> 00:34:39.579 Step one and step two are same as in case one.

479 00:34:39.579 --> 00:34:44.579 And in step one, we estimate in the missing probability.

480 00:34:45.068 --> 00:34:50.068 In step two, we estimate the data distribution.

481 00:34:53.360 --> 00:34:54.700 And then in step three,

482 00:34:54.700 --> 00:34:59.020 we get preliminary imputation estimate pf 0

483 00:35:02.690 --> 00:35:06.923 by solving this seemed a very complicated equation.

484 00:35:09.220 --> 00:35:14.220 And here there's Xl, which had two parts,

485 00:35:17.350 --> 00:35:20.640 the complete the case and on the missing part.

486 00:35:20.640 --> 00:35:24.350 The missing part is random drawn

487 00:35:24.350 --> 00:35:28.320 from this data distribution.

488 00:35:28.320 --> 00:35:29.953 We estimate from step two.

489 00:35:32.360 --> 00:35:35.290 And then the step four, okay?

490 00:35:35.290 --> 00:35:38.660 The key is that the empirical likelihood part

491 00:35:38.660 --> 00:35:43.133 where we used to compute to the weight.

492 00:35:45.791 --> 00:35:49.457 And these weights that I had, is for complete case.

493 00:35:50.360 --> 00:35:55.360 And at previous, this weight depends on three parts.

494 00:35:58.772 --> 00:36:03.772 One is missing probability, 1 is the distribution.

495 00:36:04.720 --> 00:36:09.487 Gamma previous, it depend on the preliminary as estimate

496 00:36:09.487 --> 00:36:11.490 of margin quantile.

497 00:36:11.490 --> 00:36:16.220 Now, it's related to the preliminary estimate of

498 00:36:17.892 --> 00:36:19.392 linear quantile coefficient .

499 00:36:22.359 --> 00:36:23.192 Okay?

500 00:36:23.192 --> 00:36:27.380 After we estimate these weight WI,

501 00:36:27.380 --> 00:36:32.033 then we can go to the estimating equation part, okay?

502 00:36:34.570 --> 00:36:38.463 Let's say five steps. Let's say five steps.

503 00:36:39.620 --> 00:36:43.940 As you can see you, step one, step two, step three,

504 00:36:43.940 --> 00:36:48.757 is all preexisting method we adapt trying to estimate

505 00:36:55.543 --> 00:37:00.543 the missing probability, the data distribution,

506 00:37:01.730 --> 00:37:05.200 and also impute to get a preliminary estimate

507 00:37:05.200 --> 00:37:08.300 of the parameter we are increasing.

508 00:37:08.300 --> 00:37:10.190 And then from all these,

509 00:37:10.190 --> 00:37:12.450 we pull all this information together to get

510 00:37:12.450 --> 00:37:16.403 a good weight for the compete case.

511 00:37:17.910 --> 00:37:22.910 And then the using this empirical likelihood method

512 00:37:25.110 --> 00:37:28.797 and then we adjust this complete case with the

513 00:37:30.777 --> 00:37:34.400 estimated weight to get a final estimate,

514 00:37:34.400 --> 00:37:37.113 to get the final multiple robust estimate.

515 00:37:40.990 --> 00:37:44.687 Now the case three, okay?

516 00:37:44.687 --> 00:37:48.660 In the case three, the parameter we are interested

517 00:37:48.660 --> 00:37:49.513 is still 0.

518 00:37:50.543 --> 00:37:54.780 This linear quantile regression are here.

519 00:37:54.780 --> 00:37:57.807 The scenario is the full-data vector is (Y, X).

520 00:38:01.833 --> 00:38:02.666 In this scenario, Y is missing and random, okay?

521 00:38:07.020 --> 00:38:10.130 Of course the simple complete a case analysis

522 00:38:10.130 --> 00:38:13.810 where lead to a consistent estimate,

523 00:38:13.810 --> 00:38:17.540 but it doesn't mean it will be optimal.

524 00:38:17.540 --> 00:38:21.350 Here we are trying to get a more complete educated

525 00:38:21.350 --> 00:38:24.947 but still very practical method.

526 00:38:29.500 --> 00:38:32.740 We are having some auxiliary variable.

527 00:38:32.740 --> 00:38:35.800 As this auxiliary variable,

528 00:38:35.800 --> 00:38:37.883 usually not the main study interest,

529 00:38:39.540 --> 00:38:43.221 and thus do not enter the quantile regression model.

530 00:38:43.221 --> 00:38:48.120 However, we can use it to help us to explain

531 00:38:48.120 --> 00:38:51.230 the missingness mechanism

532 00:38:51.230 --> 00:38:55.140 and to help us to build a more plausible model

533 00:38:55.140 --> 00:38:57.753 for the conditional distribution of Y.

534 00:39:00.350 --> 00:39:05.120 Now, here is the observed data.

535 00:39:06.090 --> 00:39:10.217 So, we now have an ID copies of these R, RY,

536 00:39:11.750 --> 00:39:15.943 this Y gets a missing, X is completely observed,

537 00:39:19.030 --> 00:39:21.433 and we have got auxiliary variable S.

538 00:39:23.270 --> 00:39:25.463 We have this missing and random scenario.

539 00:39:27.390 --> 00:39:32.003 We use (X, S) to denote the probability,

540 00:39:33.600 --> 00:39:38.513 and we use f(Y| X, S) to denote conditional density.

541 00:39:39.800 --> 00:39:43.340 As previous, we have multiple models

542 00:39:43.340 --> 00:39:45.750 for missing probability,

543 00:39:45.750 --> 00:39:49.873 and we have multiple models for data distribution.

544 00:39:56.320 --> 00:39:59.830 And then once again, we have the all five steps.

545 00:39:59.830 --> 00:40:03.033 The first step, we modeled the missing probability.

546 00:40:04.699 --> 00:40:09.260 And here we have this additional auxiliary variable.

547 00:40:10.180 --> 00:40:14.360 The second step, we model the data distribution.

548 00:40:14.360 --> 00:40:17.170 Again, we have this auxiliary variable.

549 00:40:17.170 --> 00:40:18.170 And then step three,

550 00:40:18.170 --> 00:40:20.689 we get a preliminary estimate on

551 00:40:20.689 --> 00:40:23.106 using this imputation method.

552 00:40:24.039 --> 00:40:28.292 We have our preliminary estimate of the parameter

553 00:40:28.292 --> 00:40:29.520 we are interested in,

554 00:40:29.520 --> 00:40:33.120 which is a linear regression coefficient here.

555 00:40:35.660 --> 00:40:39.210 And then after the preparation of step one,

556 00:40:39.210 --> 00:40:40.520 step two, and step three,

557 00:40:40.520 --> 00:40:44.303 we finally be able to estimate our weight, okay?

558 00:40:46.444 --> 00:40:48.743 Our weight is for complete case.

559 00:40:49.580 --> 00:40:51.890 And from the formula here,

560 00:40:51.890 --> 00:40:55.370 you can tell why I put this scenario as scenario three

561 00:40:55.370 --> 00:40:57.723 because it got more and more complicated.

562 00:40:58.610 --> 00:41:02.140 Although the weight still depends on three parts,

563 00:41:02.140 --> 00:41:04.504 related to the first three step.

564 00:41:04.504 --> 00:41:08.070 The missing probability related to this alpha,

565 00:41:08.070 --> 00:41:11.500 the data distribution related to this gamma,

566 00:41:11.500 --> 00:41:16.500 and the preliminary estimate made by using the imputation

567 00:41:19.140 --> 00:41:20.893 in step three.

568 00:41:24.850 --> 00:41:27.500 And once we get the weight through

569 00:41:27.500 --> 00:41:29.690 this empirical likelihood method,

570 00:41:29.690 --> 00:41:34.420 we then put it into this estimating equation.

571 00:41:34.420 --> 00:41:38.790 Adjusted by this weight, we can get our proposed estimator

572 00:41:38.790 --> 00:41:40.720 as multiple robust estimator of

573 00:41:40.720 --> 00:41:43.393 the linear regression coefficient.

574 00:41:47.850 --> 00:41:48.683 Okay.

575 00:41:49.675 --> 00:41:50.510 (coughs)

576 00:41:50.510 --> 00:41:55.180 Our method all framework in general,

577 00:41:55.180 --> 00:41:58.288 these five sets, the key thing is step four

578 00:41:58.288 --> 00:42:01.883 is empirical likelihood method to estimate the weight.

579 00:42:03.300 --> 00:42:05.531 I'll estimate his probability

580 00:42:05.531 --> 00:42:06.364 and we will estimate our framework in these three scenarios.

581 00:42:12.620 --> 00:42:14.890 Of course there are some other scenarios,

582 00:42:14.890 --> 00:42:19.890 and you can easily adapt to these five steps.

583 00:42:20.270 --> 00:42:23.280 Now, let's look at some theoretical proprietary.

584 00:42:23.280 --> 00:42:28.280 Why we propose these seemingly complicated five steps.

585 00:42:30.130 --> 00:42:35.130 We first look at the case one. There are two parts.

586 00:42:35.830 --> 00:42:40.486 The first theorem is about this consistence.

587 00:42:40.486 --> 00:42:44.363 The second theorem is about asymptotic normality, okay?

588 00:42:45.800 --> 00:42:49.190 So, under certain conditions, if...

589 00:42:50.880 --> 00:42:53.430 Remember we have two sets of models.

590 00:42:53.430 --> 00:42:57.200 One set of model, we modeled the probability.

591 00:42:57.200 --> 00:43:02.200 The other set of model, we modeled the data distribution.

592 00:43:02.200 --> 00:43:06.610 So if either one from the model

593 00:43:06.610 --> 00:43:11.160 of modeling missingness probability

594 00:43:12.090 --> 00:43:15.440 or the model set model the data distribution,

595 00:43:15.440 --> 00:43:20.193 if either one is correctly specified, Okay?

596 00:43:21.110 --> 00:43:24.013 Then, our estimate will be consistent.

597 00:43:25.604 --> 00:43:27.850 Our estimate it well be consistent.

598 00:43:27.850 --> 00:43:32.850 So, all proposed method allow you to make mistakes, okay?

599 00:43:36.770 --> 00:43:41.770 But you at least make one good right decision,

600 00:43:43.930 --> 00:43:48.660 then you get a consistent result, okay?

601 00:43:48.660 --> 00:43:51.710 Of course if you make all the bad decisions,

602 00:43:51.710 --> 00:43:54.193 you didn't choose any track modeling,

603 00:43:55.170 --> 00:43:59.030 these two sets of model, then you probably won't be able

604 00:43:59.030 --> 00:44:00.614 to get that consistent result.

605 00:44:00.614 --> 00:44:01.447 Right?

606 00:44:03.990 --> 00:44:06.930 And then the second theorem is about

607 00:44:06.930 --> 00:44:09.330 the asymptotic normality.

608 00:44:09.330 --> 00:44:14.270 Under certain conditions, the model estimate

609 00:44:16.580 --> 00:44:19.804 some multiple robust estimate on the marginal quantile

610 00:44:19.804 --> 00:44:23.124 where I have asymptotic normal distribution

611 00:44:23.124 --> 00:44:27.547 with mean zero and variates here

612 00:44:27.547 --> 00:44:30.348 is related to this variable.

613 00:44:30.348 --> 00:44:35.348 Variates is related to this data one random variable.

614 00:44:37.614 --> 00:44:42.614 And as you can see these variates of data one

615 00:44:46.421 --> 00:44:49.703 actually coming from these three parts,

616 00:44:49.703 --> 00:44:52.767 the estimate of the missingness probability,

617 00:44:52.767 --> 00:44:55.905 the estimate of these data distribution,

618 00:44:55.905 --> 00:44:59.072 and also the imputation process, okay?

619 00:45:00.105 --> 00:45:02.345 That's for case one.

620 00:45:02.345 --> 00:45:06.512 Similarly for case two, we have these two theorem.

621 00:45:08.081 --> 00:45:09.414 Y is consistent.

622 00:45:10.558 --> 00:45:13.875 And as long as the one model is correctly specified,

623 00:45:13.875 --> 00:45:16.810 we would have this consistency.

624 00:45:16.810 --> 00:45:19.727 And then this asymptotic normality,

625 00:45:20.603 --> 00:45:23.373 we would have asymptotic normal distribution.

626 00:45:23.373 --> 00:45:28.069 And also the variates, they're two, as you can see.

627 00:45:28.069 --> 00:45:31.069 The two is ready to first three step

628 00:45:31.960 --> 00:45:35.460 to estimate the different component, okay?

629 00:45:38.469 --> 00:45:41.219 And then case three, two theorem.

630 00:45:43.171 --> 00:45:47.016 Consistency, we need at least one model.

631 00:45:47.016 --> 00:45:50.478 As long as one model is correctly specified,

632 00:45:50.478 --> 00:45:52.896 we have a consistent result.

633 00:45:52.896 --> 00:45:55.507 And we have this asymptotic normalcy

634 00:45:55.507 --> 00:45:59.674 and the variates come from their three part. Okay?

635 00:46:02.143 --> 00:46:07.070 As you can see, this is a very complicated formula.

636 00:46:07.070 --> 00:46:09.775 It's a model getting more and more complicated.

637 00:46:09.775 --> 00:46:14.775 And also, if you see that you can compound the variates

638 00:46:14.874 --> 00:46:19.707 of the three to the situation with complete case analysis.

639 00:46:21.222 --> 00:46:22.630 Because for complete case analysis,

640 00:46:22.630 --> 00:46:27.630 we also get the consistent result, but like I said,

641 00:46:27.710 --> 00:46:30.240 it doesn't mean the variates would be optimal.

642 00:46:30.240 --> 00:46:34.337 And here, we actually can verify the variates of the three

643 00:46:34.337 --> 00:46:39.337 will be smaller if our model are correctly specified, okay?

644 00:46:42.530 --> 00:46:47.223 Let's say theoretical propriety.

645 00:46:48.650 --> 00:46:53.243 Now, let's look at some simulation, okay?

646 00:46:54.280 --> 00:46:57.810 We did simulation for each scenario,

647 00:46:57.810 --> 00:47:01.963 but due to the timely meet, I will only present two.

648 00:47:03.170 --> 00:47:05.170 Let's look at the second scenario.

649 00:47:05.170 --> 00:47:08.860 In the second scenario, we have four here.

650 00:47:08.860 --> 00:47:12.040 We have X1 follow exponential distribution X2

651 00:47:12.980 --> 00:47:15.563 is a normal distribution.

652 00:47:15.563 --> 00:47:20.090 And so Y is discrete, one is continuous, okay?

653 00:47:20.090 --> 00:47:24.162 The model is the simple linear model

654 00:47:24.162 --> 00:47:28.000 and the error distribution Y,

655 00:47:28.000 --> 00:47:31.870 as you can see, is heteroscedastic.

656 00:47:31.870 --> 00:47:36.333 Because of these error distribution, it's reduced to X1.

657 00:47:38.050 --> 00:47:41.760 The missing mechanism for X2,

658 00:47:41.760 --> 00:47:46.630 in the second scenario, we have a part of X2 is missing is

659 00:47:46.630 --> 00:47:49.852 through this logistic regression, okay?

660 00:47:49.852 --> 00:47:54.173 Now, missingness rate is about 38%.

661 00:47:56.710 --> 00:47:59.990 Eventually, they have this conditional quantile regression,

662 00:47:59.990 --> 00:48:03.760 linear regression, they have those coefficient excess.

663 00:48:03.760 --> 00:48:08.760 This is our simulation setup is in the second scenario.

664 00:48:12.560 --> 00:48:17.503 Now, we consider two working models for , okay?

665 00:48:19.270 --> 00:48:22.563 The fist one is correct. The second one is incorrect.

666 00:48:23.560 --> 00:48:28.560 We can see there are two models for the distribution, okay?

667 00:48:32.030 --> 00:48:32.863 All right.

668 00:48:32.863 --> 00:48:34.920 This is the incorrect one

669 00:48:34.920 --> 00:48:38.403 and for the ordinary least squares regression.

670 00:48:38.403 --> 00:48:43.403 And this is correct one with title 0.25 0.75.

671 00:48:47.740 --> 00:48:51.130 We have replication, 1,000 times.

672 00:48:51.130 --> 00:48:55.360 We have some equals 500, L is 10.

673 00:48:55.360 --> 00:48:59.384 This L is really related to the first step

674 00:48:59.384 --> 00:49:00.884 of the imputation.

675 00:49:02.550 --> 00:49:03.400 Okay.

676 00:49:03.400 --> 00:49:06.523 Now, here is all our simulation result, okay?

677 00:49:09.000 --> 00:49:13.500 Although the result has to be multiplied by 100,

678 00:49:13.500 --> 00:49:15.470 as you can see Y is very large.

679 00:49:15.470 --> 00:49:20.470 And also we denote our mass as 0000, okay?

680 00:49:24.640 --> 00:49:28.310 The fist two digit represent

681 00:49:28.310 --> 00:49:31.380 the missing probability model.

682 00:49:31.380 --> 00:49:34.610 The last two is data distribution.

683 00:49:34.610 --> 00:49:36.050 For example, for IPW 1000,

684 00:49:40.030 --> 00:49:44.490 that means we only use inverse probability method.

685 00:49:44.490 --> 00:49:49.030 And the weight is estimating is based on

686 00:49:49.030 --> 00:49:51.680 this correct weight, okay?

687 00:49:51.680 --> 00:49:55.790 And for the imputation,

688 00:49:55.790 --> 00:50:00.200 that means we only use this data distribution.

689 00:50:00.200 --> 00:50:05.200 And for this IM 0010, that means we use our first model,

690 00:50:07.790 --> 00:50:12.387 which is to model the data distribution.

691 00:50:13.823 --> 00:50:17.820 This is the second model for data distribution.

692 00:50:17.820 --> 00:50:20.030 And in either case,

693 00:50:20.030 --> 00:50:22.890 is always the first one is correct model.

694 00:50:22.890 --> 00:50:24.120 The first one is correct model.

695 00:50:24.120 --> 00:50:26.450 The second one is not, okay?

696 00:50:26.450 --> 00:50:28.260 That's just from notation.

697 00:50:28.260 --> 00:50:31.030 As you can see here using IPW

698 00:50:31.030 --> 00:50:33.540 if the model is correctly specified,

699 00:50:33.540 --> 00:50:35.300 the bias is quite small

700 00:50:35.300 --> 00:50:37.810 and everything is quite good.

701 00:50:37.810 --> 00:50:42.470 However, if you miss specify the missingness probability,

702 00:50:42.470 --> 00:50:46.940 we see the estimate is quite out of control, okay?

703 00:50:46.940 --> 00:50:51.940 Let's say for IM imputation, if you specify correctly

704 00:50:53.110 --> 00:50:55.680 the data distribution, the result is good.

705 00:50:55.680 --> 00:50:57.033 If not, then it's not.

706 00:50:57.910 --> 00:50:59.140 Okay.

707 00:50:59.140 --> 00:51:03.080 Then there's multiple robust method.

708 00:51:03.080 --> 00:51:04.857 In the multiple robust method,

709 00:51:07.910 --> 00:51:12.140 we look at, for example, this one,

710 00:51:12.140 --> 00:51:14.973 we get a missing probability correctly specified,

711 00:51:14.973 --> 00:51:17.060 then we get a good result.

712 00:51:17.060 --> 00:51:21.063 If not, we get bad result as the IPW, okay?

713 00:51:22.190 --> 00:51:27.190 But anyway, if we can choose to use all these four models,

714 00:51:29.060 --> 00:51:33.170 as you can see, the result is quite good, okay?

715 00:51:33.170 --> 00:51:37.343 The taking home method for these simulation study is,

716 00:51:38.680 --> 00:51:43.680 if you have some ideas about missingness probability

717 00:51:46.740 --> 00:51:49.690 about the state of this data distribution,

718 00:51:49.690 --> 00:51:53.050 and you think, ”Okay, maybe this one is right

719 00:51:53.050 --> 00:51:56.060 or maybe this one is also right, okay?

720 00:51:56.060 --> 00:51:58.237 So on my side, just tell you,

721 00:51:58.237 --> 00:52:01.090 ”Okay, I don’t have to just put all these

722 00:52:03.770 --> 00:52:08.197 potential candidate potential model into all framework.

723 00:52:10.680 --> 00:52:13.873 Then we look at the recount.

724 00:52:16.040 --> 00:52:21.040 This one of the simulation is scenario two.

725 00:52:21.682 --> 00:52:26.610 We also have a simulation in a scenario three,

726 00:52:26.610 --> 00:52:31.610 but I will skip it here and go directly to the

727 00:52:35.370 --> 00:52:36.320 real data analysis.

728 00:52:37.690 --> 00:52:41.097 So, in this real data analysis, we look at this

729 00:52:42.690 --> 00:52:47.690 AIDS clinical Trials Group Protocol 175 or ACTG 175 data.

730 00:52:52.230 --> 00:52:57.230 In this research, we evaluate treatment with either a single

731 00:53:00.756 --> 00:53:04.783 nucleosides or through HIV-infected subject

732 00:53:04.783 --> 00:53:06.533 whose CD4 cells count

733 00:53:07.596 --> 00:53:11.429 and are from 200 to 500 per cubic millimeters.

734 00:53:14.180 --> 00:53:16.833 So, we consider to arms or treatment.

735 00:53:16.833 --> 00:53:19.000 One is standardized,

736 00:53:19.000 --> 00:53:24.000 and the other one is with three newer treatments.

737 00:53:24.000 --> 00:53:27.703 The two arms respectively,

738 00:53:28.610 --> 00:53:32.943 have about 500 and 1,600 subjects.

739 00:53:34.020 --> 00:53:35.617 Now, model we are looking at is

740 00:53:35.617 --> 00:53:38.600 the linear quantile regression model

741 00:53:38.600 --> 00:53:43.130 and with those kind of covariates inside.

742 00:53:43.130 --> 00:53:45.853 The data can be found in this package.

743 00:53:50.600 --> 00:53:55.600 Now for the data, the average subject is 35 years old,

744 00:53:57.010 --> 00:53:59.203 standard variation is about nine,

745 00:54:01.350 --> 00:54:06.350 and the variable CD4 96 is missing for approximate 37%.

746 00:54:09.933 --> 00:54:13.633 It's quite similar to simulation scenario.

747 00:54:15.510 --> 00:54:20.510 Each athlete is part of set up of simulations scenario.

748 00:54:21.840 --> 00:54:24.660 However, at baseline during the followup,

749 00:54:24.660 --> 00:54:27.580 full measurements on additional variable are correlated

750 00:54:27.580 --> 00:54:30.410 with CD4 96 are obtained.

751 00:54:30.410 --> 00:54:35.410 So this would be the missing part. We get the missing part.

752 00:54:38.730 --> 00:54:43.730 Here we assumed this CD4 96 is the missing and random.

753 00:54:46.307 --> 00:54:50.440 And we also have other baseline, for example,

754 00:54:50.440 --> 00:54:52.320 CD4 80 and CD4 20, and so on.

755 00:54:56.470 --> 00:54:59.653 we will use these as auxiliary variables.

756 00:55:01.130 --> 00:55:06.130 So, we have our third scenario

757 00:55:06.612 --> 00:55:08.862 in this real data analysis.

758 00:55:11.852 --> 00:55:14.185 And why we choose this data?

759 00:55:15.532 --> 00:55:20.115 If we look at this CD4 96, the histogram of this, okay?

760 00:55:24.044 --> 00:55:28.127 The left one is before we do it's original skill.

761 00:55:32.340 --> 00:55:36.783 The right one is after we do log transformation.

762 00:55:38.780 --> 00:55:43.267 So, as you can see, the left one is kind of truncated,

763 00:55:45.760 --> 00:55:47.453 and the right one also truncated.

764 00:55:48.650 --> 00:55:49.527 So you may debate,

765 00:55:49.527 --> 00:55:52.430 "Okay, which one I should use?

766 00:55:52.430 --> 00:55:55.713 Do I take log transformation or not?

767 00:55:58.771 --> 00:56:00.137 Or to be, or not to be."

768 00:56:03.130 --> 00:56:08.130 So that's no apparent reason to favor one of them

769 00:56:09.803 --> 00:56:11.223 for the imputation method.

770 00:56:13.320 --> 00:56:15.993 Now, what do we do?

771 00:56:17.170 --> 00:56:19.370 In our proposed method,

772 00:56:19.370 --> 00:56:24.370 we can put all these two models in our framework, okay?

773 00:56:25.570 --> 00:56:28.173 We don't need to make the choice.

774 00:56:29.160 --> 00:56:31.120 And because no apparent reason,

775 00:56:31.120 --> 00:56:33.060 we take a log, or not take log.

776 00:56:33.060 --> 00:56:37.700 Now, let's put the two together into our model, okay?

777 00:56:37.700 --> 00:56:42.443 So we can simultaneously accommodate both simulation.

778 00:56:44.060 --> 00:56:48.700 And then we have a eight covariates and auxiliary variable.

779 00:56:48.700 --> 00:56:51.833 Then we have this probability is modeled by

780 00:56:54.300 --> 00:56:59.163 a logistic regression containing all main effect of X and S.

781 00:57:01.621 --> 00:57:04.370 So, here is the result. Here is the result.

782 00:57:04.370 --> 00:57:09.113 This is a big table, but let me summarize these table.

783 00:57:10.120 --> 00:57:10.990 Okay.

784 00:57:10.990 --> 00:57:15.043 They three newer treatment, significantly slow the progress.

785 00:57:15.941 --> 00:57:19.410 Our proposed method and the IPW method,

786 00:57:19.410 --> 00:57:22.770 produce very similar results, okay

787 00:57:22.770 --> 00:57:25.743 And the incubation estimate,

788 00:57:26.610 --> 00:57:31.180 one failed to catch difference in the treatment

789 00:57:31.180 --> 00:57:36.180 and treatment arm effect for different quantile.

790 00:57:37.506 --> 00:57:39.728 The amputation estimator 2 gives

791 00:57:39.728 --> 00:57:41.013 an increasing estimation effect and covariance.

792 00:57:43.534 --> 00:57:47.670 In addition, the two imputation estimates

793 00:57:47.670 --> 00:57:52.670 are quite sensitive to the selection of the working models.

794 00:58:03.910 --> 00:58:05.080 Okay?

795 00:58:05.080 --> 00:58:07.480 And also, from these real data,

796 00:58:07.480 --> 00:58:10.090 we can help complete case analysis

797 00:58:11.020 --> 00:58:16.020 overestimate the treatment arm effects once again,

798 00:58:16.300 --> 00:58:21.300 so that even sometimes the compete case analysis is valid

799 00:58:23.350 --> 00:58:27.593 but there are also advantage to use our proposed method.

800 00:58:33.790 --> 00:58:38.790 All right, so here's the summary of my talk.

801 00:58:40.490 --> 00:58:44.020 We proposed a general framework for

802 00:58:44.020 --> 00:58:46.593 quantile estimation with missing data.

803 00:58:48.280 --> 00:58:51.650 And we actually applied these framework

804 00:58:51.650 --> 00:58:52.943 in different scenario.

805 00:58:55.130 --> 00:58:57.290 Now, the taking home message is,

806 00:59:00.113 --> 00:59:04.290 our proposed method or whatever robust against

807 00:59:04.290 --> 00:59:07.580 possible model misspecification.

808 00:59:07.580 --> 00:59:09.820 So, as we have two sets of model,

809 00:59:09.820 --> 00:59:11.520 one for missing probability

810 00:59:11.520 --> 00:59:13.583 and one is for data distribution.

811 00:59:14.470 --> 00:59:16.610 As long as one model is correct,

812 00:59:16.610 --> 00:59:19.090 then we will get good result.

813 00:59:19.090 --> 00:59:21.750 And also our method can be easily to be generalized

814 00:59:23.132 --> 00:59:24.993 to many other scenario.

815 00:59:26.310 --> 00:59:31.310 And I think that's all of my talk,

816 00:59:32.170 --> 00:59:33.633 and thank you.

817 00:59:35.830 --> 00:59:36.663 - All right.

818 00:59:36.663 --> 00:59:39.460 Thank you, Linglong. This was very interesting.

819 00:59:39.460 --> 00:59:42.700 I think we're almost out of time, so if there's

820 00:59:42.700 --> 00:59:44.640 we have time probably for one question.

821 00:59:44.640 --> 00:59:46.340 So if there's any, if not

822 00:59:47.960 --> 00:59:49.810 Let's see if there are any questions.

823 00:59:51.760 --> 00:59:54.883 Feel free to write in the chat box or on cells.

824 01:00:12.442 --> 01:00:13.275 Okay.

825 01:00:13.275 --> 01:00:14.420 Just gonna ask one question

826 01:00:14.420 --> 01:00:16.740 and then I think I'm gonna ask all the questions

827 01:00:16.740 --> 01:00:17.573 when we meet.

828 01:00:19.166 --> 01:00:20.110 Just a quick question.

829 01:00:20.110 --> 01:00:24.400 Do you know why the complete case analysis have

830 01:00:24.400 --> 01:00:26.810 overestimation rather than underestimation?

831 01:00:26.810 --> 01:00:30.073 Like, do you have a feeling why that's the case and what?

832 01:00:33.230 --> 01:00:35.503 - Well, I don't know. No.

833 01:00:38.900 --> 01:00:39.733 - Yeah.

834 01:00:39.733 --> 01:00:42.290 I believe it will be interesting to see what cases,

835 01:00:42.290 --> 01:00:45.212 like what are the conditions for overestimation

836 01:00:45.212 --> 01:00:48.130 or underestimation for complete case analysis, I guess.

837 01:00:48.130 --> 01:00:52.280 I guess, it must depend on the data distribution

838 01:00:52.280 --> 01:00:56.320 and the missingness mechanism that's been made.

839 01:00:56.320 --> 01:00:59.480 But I'm not sure one.

840 01:00:59.480 --> 01:01:00.910 - I agree with you.

841 01:01:00.910 --> 01:01:04.790 The reason I would answer I don't know,

842 01:01:04.790 --> 01:01:09.790 because it's really hard to know how the data is miss.

843 01:01:10.990 --> 01:01:13.470 Although we assume it's missing at runtime.

844 01:01:13.470 --> 01:01:14.303 - Yeah.

845 01:01:14.303 --> 01:01:15.683 - But, who knows the reality?

846 01:01:17.190 --> 01:01:19.470 - Right. Yeah, right.

847 01:01:19.470 --> 01:01:21.930 I guess, under your assumption of missing at random,

848 01:01:21.930 --> 01:01:26.530 then I guess there could be conditions for underestimation

849 01:01:26.530 --> 01:01:29.893 or overestimation under the assumption of where MI.

850 01:01:30.860 --> 01:01:32.290 But, I don't know.

851 01:01:32.290 --> 01:01:35.702 I was wondering if people have derived those or not.

852 01:01:35.702 --> 01:01:37.410 (laughs)

853 01:01:37.410 --> 01:01:39.664 They could be future work, right?

854 01:01:39.664 --> 01:01:41.500 (laughs)

855 01:01:41.500 --> 01:01:42.889 All right.

856 01:01:42.889 --> 01:01:44.233 Linglong, thank you.

857 01:01:44.233 --> 01:01:47.120 I'll see you in an hour for a one on one meetings,

858 01:01:47.120 --> 01:01:50.920 and I know other students and maybe faculty have

859 01:01:50.920 --> 01:01:52.560 signed up for it to meet with you.

860 01:01:52.560 --> 01:01:55.040 So, thank you very much.

861 01:01:55.040 --> 01:01:56.558 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.