Estimating Heterogenous Treatment Effects in Randomized Control Trials

by Christopher Adams

Discussed by me (Salvador Navarro)
University of Western Ontario, Department of Economics

2014 Annual Health Econometrics Workshop

What I'll do

- Having never been graced with an invitation from NBER, I have never participated in this kind of conference. So I'll wing it.
- The paper is way to contrived for me to follow it as written.
- So I'll rewrite it as I would have written a first skeleton myself so I knew what I was talking about.
- This is a comment on my cognitive limitations, not on the paper itself, but more later on writting.

What I'll do

- Having never been graced with an invitation from NBER, I have never participated in this kind of conference. So I'll wing it.
- The paper is way to contrived for me to follow it as written.
- So I'll rewrite it as I would have written a first skeleton myself so I knew what I was talking about.
- This is a comment on my cognitive limitations, not on the paper itself, but more later on writting.

What I'll do

- Having never been graced with an invitation from NBER, I have never participated in this kind of conference. So I'll wing it.
- The paper is way to contrived for me to follow it as written.
- So I'll rewrite it as I would have written a first skeleton myself so I knew what I was talking about.
- This is a comment on my cognitive limitations, not on the paper itself, but more later on writting.

What I'll do

- Having never been graced with an invitation from NBER, I have never participated in this kind of conference. So I'll wing it.
- The paper is way to contrived for me to follow it as written.
- So I'll rewrite it as I would have written a first skeleton myself so I knew what I was talking about.
- This is a comment on my cognitive limitations, not on the paper itself, but more later on writting.

- Think of standard binary treatment problem where we have a treatment indicator X and two outcomes Y_1, Y_0.
- This, for reasons unknown, is not what the author does which leads to confusion (on my part) later.
- Regardless, paper makes the point that once we go beyond the nice linear operators of the mean, taking differences may not be very useful, e.g.,

$$Q(Y_1) - Q(Y_0)$$

- This is debatable, although I agree.

- Think of standard binary treatment problem where we have a treatment indicator X and two outcomes Y_1, Y_0.
- This, for reasons unknown, is not what the author does which leads to confusion (on my part) later.
- Regardless, paper makes the point that once we go beyond the nice linear operators of the mean, taking differences may not be very useful, e.g.,

$$Q(Y_1) - Q(Y_0)$$

- This is debatable, although I agree.

- Think of standard binary treatment problem where we have a treatment indicator X and two outcomes Y_1, Y_0.
- This, for reasons unknown, is not what the author does which leads to confusion (on my part) later.
- Regardless, paper makes the point that once we go beyond the nice linear operators of the mean, taking differences may not be very useful, e.g.,

$$Q(Y_1) - Q(Y_0)$$

- This is debatable, although I agree.

- Think of standard binary treatment problem where we have a treatment indicator X and two outcomes Y_1, Y_0.
- This, for reasons unknown, is not what the author does which leads to confusion (on my part) later.
- Regardless, paper makes the point that once we go beyond the nice linear operators of the mean, taking differences may not be very useful, e.g.,

$$Q(Y_1) - Q(Y_0)$$

- This is debatable, although I agree.

- Instead lets just postulate we are interested on

$$F(Y_1 - Y_0)$$

- That is, the distribution of the difference (i.e., of the treatment effect) not the difference of the distributions for whatever reason (very good reasons to be interested in it).

- This is the point of the whole paper (or should be) and why it is called "heterogenous" treatment effects. Needs to be clarified.

- One way is to get the joint

$$F(Y_1, Y_0)$$

- Instead lets just postulate we are interested on

$$F(Y_1 - Y_0)$$

- That is, the distribution of the difference (i.e., of the treatment effect) not the difference of the distributions for whatever reason (very good reasons to be interested in it).

- This is the point of the whole paper (or should be) and why it is called "heterogenous" treatment effects. Needs to be clarified.

- One way is to get the joint

$$F(Y_1, Y_0)$$

- Instead lets just postulate we are interested on

$$F(Y_1 - Y_0)$$

- That is, the distribution of the difference (i.e., of the treatment effect) not the difference of the distributions for whatever reason (very good reasons to be interested in it).

- This is the point of the whole paper (or should be) and why it is called "heterogenous" treatment effects. Needs to be clarified.

- One way is to get the joint

$$F(Y_1, Y_0)$$

- Instead lets just postulate we are interested on

$$F\left(Y_1 - Y_0\right)$$

- That is, the distribution of the difference (i.e., of the treatment effect) not the difference of the distributions for whatever reason (very good reasons to be interested in it).

- This is the point of the whole paper (or should be) and why it is called "heterogenous" treatment effects. Needs to be clarified.

- One way is to get the joint

$$F\left(Y_1, Y_0\right)$$

Missing Data and other stuff we all know

- As we all know by now that is an impossible mission, EVEN if we have a perfect RCT since we only get marginals

$$F(Y_1|X=1,R=1) = F(Y|R=1)$$
$$F(Y_0|X=0,R=0) = F(Y|R=0)$$

- Author makes the confusing claim that an RCT gives us

$$F(Y|X=1)$$
$$F(Y|X=0)$$

- This is really weird as it looks like he is claiming that a RCT gives us the selected marginal distributions.
- Of course, it all depends on how one defines X but why not use standard notation. This is where using the potential outcomes notation is helpful, it helps to distinguish

$$F(Y_j) \text{ from } F(Y_j|X=x)$$

- As we all know by now that is an impossible mission, EVEN if we have a perfect RCT since we only get marginals

$$F(Y_1|X=1,R=1)=F(Y|R=1)$$
$$F(Y_0|X=0,R=0)=F(Y|R=0)$$

- Author makes the confusing claim that an RCT gives us

$$F(Y|X=1)$$
$$F(Y|X=0)$$

- This is really weird as it looks like he is claiming that a RCT gives us the selected marginal distributions.
- Of course, it all depends on how one defines X but why not use standard notation. This is where using the potential outcomes notation is helpful, it helps to distinguish

$$F(Y_j) \text{ from } F(Y_j|X=x)$$

Missing Data and other stuff we all know

- As we all know by now that is an impossible mission, EVEN if we have a perfect RCT since we only get marginals

$$F(Y_1|X=1,R=1) = F(Y|R=1)$$
$$F(Y_0|X=0,R=0) = F(Y|R=0)$$

- Author makes the confusing claim that an RCT gives us

$$F(Y|X=1)$$
$$F(Y|X=0)$$

- This is really weird as it looks like he is claiming that a RCT gives us the selected marginal distributions.
- Of course, it all depends on how one defines X but why not use standard notation. This is where using the potential outcomes notation is helpful, it helps to distinguish

$$F(Y_j) \text{ from } F(Y_j|X=x)$$

Missing Data and other stuff we all know

- As we all know by now that is an impossible mission, EVEN if we have a perfect RCT since we only get marginals

$$F(Y_1|X=1, R=1) = F(Y|R=1)$$
$$F(Y_0|X=0, R=0) = F(Y|R=0)$$

- Author makes the confusing claim that an RCT gives us

$$F(Y|X=1)$$
$$F(Y|X=0)$$

- This is really weird as it looks like he is claiming that a RCT gives us the selected marginal distributions.
- Of course, it all depends on how one defines X but why not use standard notation. This is where using the potential outcomes notation is helpful, it helps to distinguish

$$F(Y_j) \text{ from } F(Y_j|X=x)$$

- Here things get dicey. He argues about getting the joint, but as far as I can tell that is not what he did.
- What he did is still pretty interesting though, just not what he seems to be arguing about in the introduction.
- Chris proposed to think of the problem as a mixture problem with multiple (but less than "usual") measures.
- A little tricky as both Bonhomme & Robin and Cooley, Navarro & Takahashi have shown that under conditions two measures are enough (for a more limite case though) which is what he is going to show.

- Here things get dicey. He argues about getting the joint, but as far as I can tell that is not what he did.
- What he did is still pretty interesting though, just not what he seems to be arguing about in the introduction.
- Chris proposed to think of the problem as a mixture problem with multiple (but less than "usual") measures.
- A little tricky as both Bonhomme & Robin and Cooley, Navarro & Takahashi have shown that under conditions two measures are enough (for a more limite case though) which is what he is going to show.

- Here things get dicey. He argues about getting the joint, but as far as I can tell that is not what he did.
- What he did is still pretty interesting though, just not what he seems to be arguing about in the introduction.
- Chris proposed to think of the problem as a mixture problem with multiple (but less than "usual") measures.
- A little tricky as both Bonhomme & Robin and Cooley, Navarro & Takahashi have shown that under conditions two measures are enough (for a more limite case though) which is what he is going to show.

- Here things get dicey. He argues about getting the joint, but as far as I can tell that is not what he did.
- What he did is still pretty interesting though, just not what he seems to be arguing about in the introduction.
- Chris proposed to think of the problem as a mixture problem with multiple (but less than "usual") measures.
- A little tricky as both Bonhomme & Robin and Cooley, Navarro & Takahashi have shown that under conditions two measures are enough (for a more limite case though) which is what he is going to show.

Key result

- Forget potential outcomes for a second and just think about the following problem: two variables that are correlated Y, S with the key assumption that, conditional on some other UNOBSERVABLE variable U

$$Y \perp\!\!\!\perp S | U$$

- If this is true then

$$P(Y < y, S < s) = \Sigma_u \pi(u) F_{Y|U}(y) G_{S|U}(s)$$

- Important assumption not mentioned (as an assumption and only mentioned later in the paper): discreteness. Should be made explicit.
- This may sound trivial but most results on mixtures are discontinuous at the limit (and I suspect this one is too)

Key result

- Forget potential outcomes for a second and just think about the following problem: two variables that are correlated Y, S with the key assumption that, conditional on some other UNOBSERVABLE variable U

$$Y \perp\!\!\!\perp S | U$$

- If this is true then

$$P(Y < y, S < s) = \Sigma_u \pi(u) F_{Y|U}(y) G_{S|U}(s)$$

- Important assumption not mentioned (as an assumption and only mentioned later in the paper): discreteness. Should be made explicit.
- This may sound trivial but most results on mixtures are discontinuous at the limit (and I suspect this one is too)

Key result

- Forget potential outcomes for a second and just think about the following problem: two variables that are correlated Y, S with the key assumption that, conditional on some other UNOBSERVABLE variable U

$$Y \perp\!\!\!\perp S|U$$

- If this is true then

$$P(Y < y, S < s) = \Sigma_u \pi(u) F_{Y|U}(y) G_{S|U}(s)$$

- Important assumption not mentioned (as an assumption and only mentioned later in the paper): discreteness. Should be made explicit.
- This may sound trivial but most results on mixtures are discontinuous at the limit (and I suspect this one is too)

Key result

- Forget potential outcomes for a second and just think about the following problem: two variables that are correlated Y, S with the key assumption that, conditional on some other UNOBSERVABLE variable U

$$Y \perp\!\!\!\perp S | U$$

- If this is true then

$$P(Y < y, S < s) = \Sigma_u \pi(u) F_{Y|U}(y) G_{S|U}(s)$$

- Important assumption not mentioned (as an assumption and only mentioned later in the paper): discreteness. Should be made explicit.
- This may sound trivial but most results on mixtures are discontinuous at the limit (and I suspect this one is too)

Key results

- In matrix notation (more discreetenes)

$$\underset{I\times J}{P} = \left[\underset{I\times K}{F}\underset{K\times K}{D_\pi}\right]\underset{K\times J}{G'}$$

- Assumption (needs to be made explicit) $I \geq K$, $J \geq K$. What this means is that we have at least as many states that Y (I) and S (J) can take as there are on the unobservable U (K).
- We know that it is also true that

$$P = \underset{I\times K}{W}\underset{K\times J}{H}$$

- But this factorization may not be unique.

- In matrix notation (more discreetenes)

$$\underset{I\times J}{P} = \left[\underset{I\times K}{F}\underset{K\times K}{D_\pi}\right]\underset{K\times J}{G'}$$

- Assumption (needs to be made explicit) $I \geq K$, $J \geq K$. What this means is that we have at least as many states that Y (I) and S (J) can take as there are on the unobservable U (K).
- We know that it is also true that

$$P = \underset{I\times K}{W}\underset{K\times J}{H}$$

- But this factorization may not be unique.

Key results

- In matrix notation (more discreetenes)

$$\underset{I\times J}{P} = \left[\underset{I\times K}{F}\underset{K\times K}{D_\pi}\right]\underset{K\times J}{G'}$$

- Assumption (needs to be made explicit) $I \geq K$, $J \geq K$. What this means is that we have at least as many states that Y (I) and S (J) can take as there are on the unobservable U (K).
- We know that it is also true that

$$P = \underset{I\times K}{W}\underset{K\times J}{H}$$

- But this factorization may not be unique.

Key results

- In matrix notation (more discreetenes)

$$\underset{I\times J}{P} = \left[\underset{I\times K}{F}\,\underset{K\times K}{D_\pi}\right]\underset{K\times J}{G'}$$

- Assumption (needs to be made explicit) $I \geq K$, $J \geq K$. What this means is that we have at least as many states that Y (I) and S (J) can take as there are on the unobservable U (K).
- We know that it is also true that

$$P = \underset{I\times K}{W}\,\underset{K\times J}{H}$$

- But this factorization may not be unique.

- Under “mild” conditions (non-ngeativity and sparseness) it is unique (up to permutation, i.e., relabeling)
- Sparseness (zeros) is the key one: as you will see it will have the flavor of identification at infinity
- We can be “close” to unique if we are “close” to sparse

- Under "mild" conditions (non-ngeativity and sparseness) it is unique (up to permutation, i.e., relabeling)
- Sparseness (zeros) is the key one: as you will see it will have the flavor of identification at infinity
- We can be "close" to unique if we are "close" to sparse

- Under "mild" conditions (non-ngeativity and sparseness) it is unique (up to permutation, i.e., relabeling)
- Sparseness (zeros) is the key one: as you will see it will have the flavor of identification at infinity
- We can be "close" to unique if we are "close" to sparse

Theorem 2

- With this we can state theorem 2 (the main theorem): if blah, blah, blah then
$$P = FD_\pi G'$$
is unique up to relabeling.
- What is the if blah, blah? A bunch of uninteresting technical stuff and an interesting one: for each type (K) we want there to exist states (outcomes) that are not possible but highely likely for the other types.
- It seems to me that this is "like" identification at infinity. If I had regressors I would like to vary them so that I know that a type will never visit a state.
- In schooling this is like saying that as Z goes to inifinity you will never be just a high school graduate.

Theorem 2

- With this we can state theorem 2 (the main theorem): if blah, blah, blah then
$$P = FD_\pi G'$$
is unique up to relabeling.
- What is the if blah, blah? A bunch of uninteresting technical stuff and an interesting one: for each type (K) we want there to exist states (outcomes) that are not possible but highely likely for the other types.
- It seems to me that this is "like" identification at infinity. If I had regressors I would like to vary them so that I know that a type will never visit a state.
- In schooling this is like saying that as Z goes to inifinity you will never be just a high school graduate.

Theorem 2

- With this we can state theorem 2 (the main theorem): if blah, blah, blah then
$$P = FD_\pi G'$$
is unique up to relabeling.
- What is the if blah, blah? A bunch of uninteresting technical stuff and an interesting one: for each type (K) we want there to exist states (outcomes) that are not possible but highely likely for the other types.
- It seems to me that this is "like" identification at infinity. If I had regressors I would like to vary them so that I know that a type will never visit a state.
- In schooling this is like saying that as Z goes to inifinity you will never be just a high school graduate.

Theorem 2

- With this we can state theorem 2 (the main theorem): if blah, blah, blah then
$$P = FD_\pi G'$$
is unique up to relabeling.
- What is the if blah, blah? A bunch of uninteresting technical stuff and an interesting one: for each type (K) we want there to exist states (outcomes) that are not possible but highely likely for the other types.
- It seems to me that this is "like" identification at infinity. If I had regressors I would like to vary them so that I know that a type will never visit a state.
- In schooling this is like saying that as Z goes to inifinity you will never be just a high school graduate.

- It is not quite that as this are joint states (i.e., Y, S states) and there are no regressors.
- I would have liked to see an example of when this is likely to happen.
- It seems to me that this is more likely if $I, J \gg K$.
- If $K = 2$ say, then I want there two be a pair of Y, S that $k = 1$ will not visit (very young people will not die immediately if they are type 1 but may if they are type 2) and also for $k = 2$ (very old people will not survive long if they are type 2 but type 1's will).
- Regressors?

- It is not quite that as this are joint states (i.e., Y, S states) and there are no regressors.
- I would have liked to see an example of when this is likely to happen.
- It seems to me that this is more likely if $I, J \gg K$.
- If $K = 2$ say, then I want there two be a pair of Y, S that $k = 1$ will not visit (very young people will not die immediately if they are type 1 but may if they are type 2) and also for $k = 2$ (very old people will not survive long if they are type 2 but type 1's will).
- Regressors?

- It is not quite that as this are joint states (i.e., Y, S states) and there are no regressors.
- I would have liked to see an example of when this is likely to happen.
- It seems to me that this is more likely if $I, J \gg K$.
- If $K = 2$ say, then I want there two be a pair of Y, S that $k = 1$ will not visit (very young people will not die immediately if they are type 1 but may if they are type 2) and also for $k = 2$ (very old people will not survive long if they are type 2 but type 1's will).
- Regressors?

- It is not quite that as this are joint states (i.e., Y, S states) and there are no regressors.
- I would have liked to see an example of when this is likely to happen.
- It seems to me that this is more likely if $I, J \gg K$.
- If $K = 2$ say, then I want there two be a pair of Y, S that $k = 1$ will not visit (very young people will not die immediately if they are type 1 but may if they are type 2) and also for $k = 2$ (very old people will not survive long if they are type 2 but type 1's will).
- Regressors?

- It is not quite that as this are joint states (i.e., Y, S states) and there are no regressors.
- I would have liked to see an example of when this is likely to happen.
- It seems to me that this is more likely if $I, J \gg K$.
- If $K = 2$ say, then I want there two be a pair of Y, S that $k = 1$ will not visit (very young people will not die immediately if they are type 1 but may if they are type 2) and also for $k = 2$ (very old people will not survive long if they are type 2 but type 1's will).
- Regressors?

- Back to potential outcomes. How is this helpful? For my money it is not very clear in the paper.
- I think this is what the author means to say. Since

$$P(Y_j < y) = \Sigma_u \pi(u) F_{Y_j|U}(y)$$

- Then even if we find that

$$P(Y_1 < y) - P(Y_0 < y) > 0$$

it can be that

$$P(Y_1 < y, U = u) - P(Y_0 < y, U = u) < 0$$

for a lot of people!

- Back to potential outcomes. How is this helpful? For my money it is not very clear in the paper.
- I think this is what the author means to say. Since

$$P(Y_j < y) = \Sigma_u \pi(u) F_{Y_j|U}(y)$$

- Then even if we find that

$$P(Y_1 < y) - P(Y_0 < y) > 0$$

it can be that

$$P(Y_1 < y, U = u) - P(Y_0 < y, U = u) < 0$$

for a lot of people!

- Back to potential outcomes. How is this helpful? For my money it is not very clear in the paper.
- I think this is what the author means to say. Since

$$P(Y_j < y) = \Sigma_u \pi(u) F_{Y_j|U}(y)$$

- Then even if we find that

$$P(Y_1 < y) - P(Y_0 < y) > 0$$

it can be that

$$P(Y_1 < y, U = u) - P(Y_0 < y, U = u) < 0$$

for a lot of people!

- So, I think, this is what he is after. Looking at the difference conditional on type (i.e., values of U) as a better description of what is going on "groupwise"
- This is not the same as what the argument in the introduction seems to be which is to go for

$$F(Y_1 - Y_0)$$

- But it still is an interesting object. i.e., the distribution of individual effects
- Assume that this is what he does (I apologize for not contacting him)

- So, I think, this is what he is after. Looking at the difference conditional on type (i.e., values of U) as a better description of what is going on "groupwise"
- This is not the same as what the argument in the introduction seems to be which is to go for

$$F(Y_1 - Y_0)$$

- But it still is an interesting object. i.e., the distribution of individual effects
- Assume that this is what he does (I apologize for not contacting him)

- So, I think, this is what he is after. Looking at the difference conditional on type (i.e., values of U) as a better description of what is going on "groupwise"
- This is not the same as what the argument in the introduction seems to be which is to go for

$$F(Y_1 - Y_0)$$

- But it still is an interesting object. i.e., the distribution of individual effects
- Assume that this is what he does (I apologize for not contacting him)

- Then follows that we can repeat the exercise for

$$F(Y_j < y, S < s) = \Sigma_u \pi(u) F_{Y_j|U}(y) G_{S|U}(s)$$

- We can obtain the distribution of the mixture and calculate

$$P(Y_j < y) = \Sigma_u \pi(u) F_{Y_j|U}(y)$$

and hence

$$P(Y_1 < y, U = u) - P(Y_0 < y, U = u) < 0$$

- Then follows that we can repeat the exercise for

$$F(Y_j < y, S < s) = \Sigma_u \pi(u) F_{Y_j|U}(y) G_{S|U}(s)$$

- We can obtain the distribution of the mixture and calculate

$$P(Y_j < y) = \Sigma_u \pi(u) F_{Y_j|U}(y)$$

and hence

$$P(Y_1 < y, U = u) - P(Y_0 < y, U = u) < 0$$

- We can do it for each Y_j and test that π and G are the same, or we can do it jointly imposing that $\pi(u)$ and $G(S)$ is the same for all j.
- Estimation is then "simply" a matter of doing (constrained) minimum distance (or GMM)
- He does a Montecarlo exercise.
- More interestingly he estimates this on Chemotherapy patients (Stage III colon cancer, incidentally what my father died of a year ago)

- We can do it for each Y_j and test that π and G are the same, or we can do it jointly imposing that $\pi(u)$ and $G(S)$ is the same for all j.
- Estimation is then "simply" a matter of doing (constrained) minimum distance (or GMM)
- He does a Montecarlo exercise.
- More interestingly he estimates this on Chemotherapy patients (Stage III colon cancer, incidentally what my father died of a year ago)

- We can do it for each Y_j and test that π and G are the same, or we can do it jointly imposing that $\pi(u)$ and $G(S)$ is the same for all j.
- Estimation is then "simply" a matter of doing (constrained) minimum distance (or GMM)
- He does a Montecarlo exercise.
- More interestingly he estimates this on Chemotherapy patients (Stage III colon cancer, incidentally what my father died of a year ago)

- We can do it for each Y_j and test that π and G are the same, or we can do it jointly imposing that $\pi(u)$ and $G(S)$ is the same for all j.
- Estimation is then "simply" a matter of doing (constrained) minimum distance (or GMM)
- He does a Montecarlo exercise.
- More interestingly he estimates this on Chemotherapy patients (Stage III colon cancer, incidentally what my father died of a year ago)

- Two treatments: “observation” after surgery, just Lev or chemotherapy (5-Fu) .
- The mean effect is that Chemotherapy reduces the risk of recurrence by 41% and the death rate by 33%.
- The author instead models it as a mixture of two types of patients with histiology and number of lymph nodes affected being the signals S.
- He actually studies three treatments (although he only talks about two at the beginning). Observation, Lev and Lev + 5Fu.

- Two treatments: “observation” after surgery, just Lev or chemotherapy (5-Fu) .
- The mean effect is that Chemotherapy reduces the risk of recurrence by 41% and the death rate by 33%.
- The author instead models it as a mixture of two types of patients with histiology and number of lymph nodes affected being the signals S.
- He actually studies three treatments (although he only talks about two at the beginning). Observation, Lev and Lev + 5Fu.

- Two treatments: "observation" after surgery, just Lev or chemotherapy (5-Fu) .
- The mean effect is that Chemotherapy reduces the risk of recurrence by 41% and the death rate by 33%.
- The author instead models it as a mixture of two types of patients with histiology and number of lymph nodes affected being the signals S.
- He actually studies three treatments (although he only talks about two at the beginning). Observation, Lev and Lev + 5Fu.

- Two treatments: “observation” after surgery, just Lev or chemotherapy (5-Fu) .
- The mean effect is that Chemotherapy reduces the risk of recurrence by 41% and the death rate by 33%.
- The author instead models it as a mixture of two types of patients with histiology and number of lymph nodes affected being the signals S.
- He actually studies three treatments (although he only talks about two at the beginning). Observation, Lev and Lev + 5Fu.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- 77% are type 1. For these types all treatments are virtually identical.
- For type 2 patients though observation is a death sentence (no one survives 4 years), 15% do with Lev and 51% do with Lev + 5-Fu.
- Little discussion about how to identify the types. Clearly we can try to via the signals we observe.
- In his case, type 2 patients are those that are likely to have more than 4 lymph nodes affected (and poorly or well differentiated tumors).
- Some mention should be made that the more signals we have the better we can predict patient type. So we can do it with just 1, but having more is better from a practical (forget indentification) view.
- Sparsity not very well satisfied though.

- Very interesting idea
- But it is not going after the joint in the sense that he seems to describe in the introduction.
- In the papers we have worked on we are talking about

$$F(Y_1, Y_0) = \int F(Y_1|U) F(Y_0|U) dF(U)$$

but that is what you have as you are interpreting this as types so $\pi(u)$ is the probability of type u regardless of $X = 1$ or $X = 0$.
- For that, I would write the model as $X_i = a_i(U_i), Y_{i1} = b_{i1}(U_i), Y_{i0} = b_{i0}(U_i), S_i = c_i(U_i)$.

- Very interesting idea
- But it is not going after the joint in the sense that he seems to describe in the introduction.
- In the papers we have worked on we are talking about

$$F(Y_1, Y_0) = \int F(Y_1|U) F(Y_0|U) dF(U)$$

but that is what you have as you are interpreting this as types so $\pi(u)$ is the probability of type u regardless of $X = 1$ or $X = 0$.
- For that, I would write the model as $X_i = a_i(U_i), Y_{i1} = b_{i1}(U_i), Y_{i0} = b_{i0}(U_i), S_i = c_i(U_i)$.

- Very interesting idea
- But it is not going after the joint in the sense that he seems to describe in the introduction.
- In the papers we have worked on we are talking about

$$F(Y_1, Y_0) = \int F(Y_1|U) F(Y_0|U) dF(U)$$

but that is what you have as you are interpreting this as types so $\pi(u)$ is the probability of type u regardless of $X = 1$ or $X = 0$.
- For that, I would write the model as $X_i = a_i(U_i), Y_{i1} = b_{i1}(U_i), Y_{i0} = b_{i0}(U_i), S_i = c_i(U_i)$.

- Very interesting idea
- But it is not going after the joint in the sense that he seems to describe in the introduction.
- In the papers we have worked on we are talking about

$$F(Y_1, Y_0) = \int F(Y_1|U) F(Y_0|U) dF(U)$$

  but that is what you have as you are interpreting this as types so $\pi(u)$ is the probability of type u regardless of $X = 1$ or $X = 0$.
- For that, I would write the model as $X_i = a_i(U_i), Y_{i1} = b_{i1}(U_i), Y_{i0} = b_{i0}(U_i), S_i = c_i(U_i)$.

- The fact that S is common is crucial, but no mention is done. It could be done without that but it would be silly as it would be purely imposing it.
- That is, we can identify something about how Y_1 and Y_0 are linked because there is a common equation between the two: S. Otherwise, we would be simply imposing that π is the same without anything linking them.
- In our work on estimation of joint distributions that role can be played by the selection equation itself. Hence one could think of doing this without RCT.
- Paper needs to be a lot more explicit about what is actually done.

- The fact that S is common is crucial, but no mention is done. It could be done without that but it would be silly as it would be purely imposing it.
- That is, we can identify something about how Y_1 and Y_0 are linked because there is a common equation between the two: S. Otherwise, we would be simply imposing that π is the same without anything linking them.
- In our work on estimation of joint distributions that role can be played by the selection equation itself. Hence one could think of doing this without RCT.
- Paper needs to be a lot more explicit about what is actually done.

- The fact that S is common is crucial, but no mention is done. It could be done without that but it would be silly as it would be purely imposing it.
- That is, we can identify something about how Y_1 and Y_0 are linked because there is a common equation between the two: S. Otherwise, we would be simply imposing that π is the same without anything linking them.
- In our work on estimation of joint distributions that role can be played by the selection equation itself. Hence one could think of doing this without RCT.
- Paper needs to be a lot more explicit about what is actually done.

- The fact that S is common is crucial, but no mention is done. It could be done without that but it would be silly as it would be purely imposing it.
- That is, we can identify something about how Y_1 and Y_0 are linked because there is a common equation between the two: S. Otherwise, we would be simply imposing that π is the same without anything linking them.
- In our work on estimation of joint distributions that role can be played by the selection equation itself. Hence one could think of doing this without RCT.
- Paper needs to be a lot more explicit about what is actually done.