Survival Analysis: An Introduction
April 26, 2018

Erin Farley: ... research and analysis to inform criminal and juvenile justice system decision making, and we are comprised of a network of researchers and practitioners, which at the core include directors and staff from the state's statistical analysis centers.

Erin Farley: It is my pleasure today to welcome, new to our webinar on survival analysis, and it will be presented by Dan O'Connell, who is a senior scientist at the Center for Drug and Health Studies and is also an assistant professor of criminal justice at the University of Delaware. He is an expert in programs evaluation, research on criminal careers, and the application of statistical methods to criminal justice issues. So, welcome.

Erin Farley: Before we go any further I just want to thank our partners at the Bureau of Justice Statistics for helping to make this webinar possible, and also cover a few logistical items. So for those of you who have been here watching our slides click along you might have seen that we do have some upcoming webinars. We do have one on May 3rd that is going to be by Matt Landon, presented by Matt Landon who is at the Washington Tech, and then we also have another one ... and you can actually register for that one right now, it's open.

Erin Farley: And then we do have another one that's on June 14th, and that actually, registration should be opening very soon. So looking forward to those. And we do have additional webinars tentatively scheduled all the way through the summer, and so we will be posting those and opening registration over the summer. So keep an eye out for those.

Erin Farley: As you may be familiar, we are audio casting this webinar through speakers, and obviously via your computer. You can, if you are having any audio problems, you can go to the menu bar at the top, click on Audio, go to Audio Conference, and then it will either have a number that you can dial in and using your phone, or you can use the speakers on your computer.

Erin Farley: If you have any audio issues throughout this session you can just, again, click on that Audio button or you can follow the link that was given to you when you registered for the conference, so there's a couple options. If you still run into any issues you can contact Jason Trask at JTrask@jrsa.org and he will try, and help you. This session is being recorded and it will be available in the next couple of days on our website, so if you'd like to review it or recommend it to some colleagues you will be able to find it in just a couple of days.

Erin Farley: Let's see ... okay, oh yes. So if anybody has any questions, as usual we do mute everybody on entry, so if you have questions please use the chat feature that's at the bottom right hand corner of your screen, and please remember to select Host, Presenter, and/or panelist so that will go out to everybody. For example, Jason Trask is the host, so if you sent a question I actually won't see it or Dan won't see it. And Dan if you don't mind, one thing we can do is if people pop up
with questions I could just sort of politely interrupt and let you know what the question is, if that sounds good to you.

Daniel O’Connell: Yeah, that's fine.

Erin Farley: Okay. Let's see. And then, helping us count. So we like to keep track of how many people are calling in to attend our webinars, and so we do know that there are situations where people use one computer, so if you are one of those if you could just put in the chat feature, your name of the person who registered and how many people are watching, that just really helps us count, and keep track of whose here, how many people are actually watching it actively.

Erin Farley: And then, let's see, did I cover everything? This session is scheduled for about one hour. We do have our poll at the end of the session, and that's just a few questions that again, just gives us feedback and helps us keep track of how we're doing and what we can do to improve. Those are our two upcoming webinars. Again, the May 3rd is open for registration and June 14th will be open soon. Okay, and so let me ... I think that is it. That's all I have to say. So Dan, I'm going to take the magic ball and pass it over to you.

Daniel O’Connell: Thank you. I was waiting for the magic ball to come. I'm Dan O'Connell, like Erin said I'm a senior scientist at the Center for Drug and Health Studies. They asked me to put together a little something to walk people through, an introduction to survival analysis. I want to focus on the key word up there, which is introduction. This is not a high end survival analysis. Matter of fact, mostly what I'm going to talk about is time and how we think about time and how we analyze time.

Daniel O’Connell: So with that, lets kind of move forward. Why are we here? Here we want to get an understanding of time. I want to talk a little bit about the data, and then I always say, start simple. Survival analysis should be the last thing you should do in your analysis. If you're doing time stuff you should work your way up to it. Then we're going to talk about survival tables, Kaplin-Meier plots, Cox Regressions, those are the two analytic techniques that we'll discuss here once we get to them.

Daniel O’Connell: So why, what is the big picture of this? We're often asked to report on programs or initiatives or different things that are happening in criminal justice. We want to get the right answers to those questions so that we can inform policy in a good way, and we want to do that in a way that we can present to general audiences.

Daniel O’Connell: So for this one in particular, we want to move from a question to data, to analysis, to recording, all right. So you should know what a survival analysis is good for. You should know a little about the required data parameters. You should know how to run them all. I'm going to be using SPSS for this, and I've got some screen shots for how to work your way through this. [R 00:06:47] will
certainly do this, task will do this, data will do this. And but for today, I'm going to run us through the SPSS interpretations. I'm going to talk about how to interpret the output in these models and then a little bit about how to explain the results to general audiences.

Daniel O’Connell: All right. So understanding time dependent questions. Not all questions are suitable for survival analyses. The key feature is always time. How long did it take for X to happen? X for us is actually, I'll just use the example that I'm going to use, which is arrested for a new crime after a particular date. So does X happen at different times for different groups? In today's example, it'll be a treatment program that people were involved in and so the question becomes, did X happen at different times for those two groups? So often we'll have not only did one group do better than the other, in terms of the full time but did people last longer essentially? Because when we're doing programming and we're talking to policy makers, one of the additional benefits of a change in policy or a change in a program is not necessarily that everybody desists and goes out and lives mad happy lives. But if you can just prevent them from getting arrested, for an additional six months or whatever the time period is, that's another way of showing that something that you're doing is working.

Daniel O’Connell: When we get to the Cox model, they're really good at predicting, well the Kaplan-Meier and the Cox models are how predicting how long it takes for X to happen, and what predicts that change in length of time. All right. So the example we're going use today is, data from a probation based treatment program. And so the two questions are, was the overlying question is, was the program effective at reducing recidivism? In this case, rearrest for a new crime and that's a zero one variable, arrested or not. Was the recidivism rate different for groups? And key for us is how long did it take them to fail? So we really want to focus in on the time variable. There's 400 people in the sample, 200 in each group. And we'll be looking at some background characteristics on race, gender, age, and age at first arrest.

Daniel O’Connell: So it's all about the data. In this case, it's all about the dates. Date are very, very particular variables. Very problematic data. Often they are widely susceptible to data entry errors and field entry errors. So you should really spend a lot of time making sure that your dates are correct, and understanding exactly what dates you should use. Administrative databases, the types of [facts 00:09:59] often work with.

Daniel O’Connell: You'll be looking at perhaps, you know let's just take a reentry variable like when somebody left a date, somebody left an institution, well, where did they go? Did they move from one institution to another? Did they move from an institution to a lower level institution? You have to figure out exactly what the dates are and usually what we want for these types of things, is the date someone became at risk. So the date at which they went to the street.
Daniel O'Connell: Often when you're doing survival analysis and things you need k types, you're eventually going to need to get to the number of days until something which will have you taking like an arrest date and subtracting out at risk date and often you'll get negative numbers on that. These can be problematic. And an indication that you need to go back and revisit your dates and making sure that those days are all right.

Daniel O'Connell: Generally, when I do this kind of analysis, the majority of time that we spent is not actually running the models, it's getting the the date data correct, 'cause if your date data isn't right, then anything that you do that follows from it, is not going to be right either. So Dates are very nice but you can't really analyze them. So you have to take dates and move them to time. In this case, time counts the number of days. We can use days, weeks, months, years, whatever. Generally we use days, so the number of days until something happens.

Daniel O'Connell: In our example, we're going to use arrest, so to figure this out, then need a number of dates. We need a number of dates. We need the start date which is the date somebody became at risk. In the current study, it was the date you were enrolled into the program 'cause this was a probation study, so the date somebody basically the date they were randomized was what we use. But they could be any at risk dates, the date they got out of prison, or jail. But basically, it's your start date. When does the clock start?

Daniel O'Connell: And what we're looking at is the amount of time to the arrest date. So how many days did it take? So those are your two dates. The problem comes because everybody in most administrative databases and pretty much any database until you manipulate it, if somebody doesn't get arrested, or experience whatever the event is, they don't have a date for that. So we create an end date which is the date the data collection ended or your cut off date, whatever that is. And we have to assign that date to a variable so that everybody has two valid dates, otherwise, the data go missing. And once the data go missing, obviously, we can't analyze it.

Daniel O'Connell: So the way we do this, I'm going to fire up my little pointer here so I can highlight somethings. Hopefully you see that little red dot, that's me. All right. So we want to combine the arrest date and the end date. So assign an end date. In this particular study, it was April 15th of 2013. That's the day the study ended in data collection stops. So that would be essentially our right-hand censored date. All right.

Daniel O'Connell: So we have this firstpost released date, that's the date of the first post release arrest. Okay, so what we need to do then, is create a new variable which we're calling in this case, alldate, which is going to start with that firstpost date and what we do then is if you're missing on that, meaning you didn't get arrested, we assign you that end date, so this new variable then has a combination of the arrest date and the end date such that, everybody has a valid date. And I will show you here how that works. Well I'll show you in a minute.
So once we do that, once we have this variable created, we can then go to SPSS and we're going to want to compute the new variable. All right. So we're going subtract the start date from the arrest date. So this is a basic compute statement in whatever software it is. In SPSS it would compute date arithmetic, date difference. So it's a computer a new variable, and go over to here and we're looking for date arithmetic and then I don't know why DATESUM is highlighted. It should be DateDiff, so that's the date difference.

And what we're doing is subtracting start date from end date, and we want to units to be days. Cool thing about SPSS when you click on one of these, it will give you a definition of exactly what its doing there. Or to send the SPSS syntax is right here for this Compute timetoar is a DateDiff so you can run it that way, or has its own, Stata has its own. These are fairly simple procedures.

Anyway, what that then gives us, here's screen shot of what we just did. Okay. So we had start date which is the date people became at risk. Then we had firstpost which is the arrest date for everybody that was arrested, that's these folks here, but those people who were not arrested we're missing on this. And what we did a couple of slides ago, was then to, if you were ... we created a new variable here, we called it alldate and if you were missing here, we assigned you the end date of the study. So everybody that was missing became 4/15/13. All right?

And then so then we have a variable where everybody has valid dates here, everybody has a valid date, here. We then did that compute statement for a new variable, moving from dates, these are all dates. This is a count variable. This is number of days to arrest and you see those people that didn't get arrested. Obviously they have a lot of dates.

But now we've got a variable that we can analyze and nobody is missing on it. So we take the step. And I always say start simple. I tell this to my students all the time. They always want to rush in and jump right to the regression of the survival analysis and I tell them, "No, no, no. Start simple and work your way up." It's kind of going back to your introductory stats class, I will you know. Forgive me for that. But I feel like I need to drill this. So we work up to frequencies, means, look for distribution, and outliers. Then we move to bivariate stuff, and then finally to survival. So very quickly, going to show you how I do these when I'm dealing with time and then we're get on to the survival stuff.

Frequencies. This is just getting to know you, getting to know your data. For this one, we had about 35% who were arrested for a new crime during the study period, which was about a year. And then we go to the descriptives. Here's our timear variables. The minimum we had somebody who made it a whooping day. And somebody made it about three years. That's actually an outlier that we reigned in and our mean was 236.
Daniel O'Connell: So doing this yourself, your starting to get a feel for your data as you move towards your survival runs which are coming later. All right. So when we ask a question, when usually when your examining a program or change in policy, something that there's something that one group got and one group didn't get, that's a basic zero one variable. But for studies like this, anybody whose done recidivism studies knows you have to have time. So what generally ... in this study what we did was we segmented it out by days. We did 90 days, 180, 270, basically core our years was our reporting. You know the metric for a standard recidivism study is three years where you would be showing, one year, two years, three years, the same rules apply here.

Daniel O'Connell: So what we want to do so we want some SPSS or whatever you're using, we need to select people who were at risk for the amount of time. So in this case, I want to look at the only if people who had 90 days at risk. Right. So from our sample of 400, were already dropping, we're down to 305 and we get our recidivism rate there. And here's the the temporary ... you do a temporary command on this, make your select and start walking through the time. Right. And so we did that and then we went to 180, 270, 365.

Daniel O'Connell: So we're already modeling time but we're modeling time at the bivari ... well actually at univariant level will go to bivariate in a minute. Couple of things you see as we go out in time as we expect, the percent recidivating goes up pretty high so that 39% recidivism figure, as people get higher out, that number gets quite high as the sample shrinks 'cause in this case, the study had an end date, so not everybody was in at the end which is handy thing for survival analysis, we'll talk about in a bit.

Daniel O'Connell: Okay. What we really want to know is the effect of condition so we move from those univariant to our bivariate and this is the same thing. The important thing here is instead of just writing crosstab on everybody, we want to run crosstabs by time at risk. I'm not going to go through all of these. So we're selecting out those who are at risk for 90 then we did that. I don't have slides for this. I'm just showing the one. And then we're looking at our condition variable which was standard probation or treatment program. In this case, I add 90 days. Our treatment people were less likely to recidivate than our standard probation people that's significant.

Daniel O'Connell: And the reason I'm showing this again, is that you really do need to start simple and start working your way up to the those survival functions. All right. So then we go to our next five variant tests which are key T-Test looking at. Now we're really starting to model time, okay? So we're looking at days to arrest here. So we're looking at the condition on that variable we created which was timetoar which includes that cut off date for everybody and we're running at just a simple T-Test and we're looking at the means days to arrest by group so the standard group, they fell out where likely their mean arrest time was 184 days. The mean arrest time or end of study date was 214 for the treatment folks. That's clearly significant down here.
So I stressed all that because again, you should not start with survival analysis. So you run those types of things to get a picture for what's going on and by the time you get here, you already know quite a bit. So one question is okay if we know all this, why do a survival analysis? Couple of reasons. Visual plot are effective tools. When we get to these Kaplan-Meier curves, you look at one of these things and especially if you've got it, well, even if you don't have it at significant effect, but when you have a significant effect, you can show a policy maker or a layperson and one look at these charts and you can really get the impact of something. If it's working, or conversely if it's not working, it's like, look these lines are the same.

But the main reason, Cox Regression especially, allow us to assess the impact of something like a program on time with something else. Cox Regression are funky because you really have two dependent variables. You have time and event. I'll talk more about that later. But we can control for other variables while we're doing this. So as with any regression, we did our univariant, our bivariate, now we want to move up to the multivariant and that's why we need these Cox Regression models up here, which are coming up here in a minute.

And importantly, all Kaplan-Meier and Cox, survival analysis uses all of the data by modeling the censored data. All right? So people are censored. This came out of ... all of this stuff came out of the medical field, you know whether you're doing studies on heart diseases and you're looking at the percentage of people who died and survived over a long periods of time, but you very rarely follow them to the end of their lives and so everybody is essentially censored. And that's true here, even doing a three year recidivism study where censoring people at the end of the three year window, we know they didn't die in three years or they didn't recidivate in three years but we don't know if they would have recidivated at all.

Anyway, I'm getting a little off. But the censor data is especially important for people who don't make it to the end of a time period. Like in our case, with a study date of folks kind of fell out, it models using all of the data so that people don't go missing by modeling either the event, or the censored date. And I'll talk about that here in a little bit.

All right. So survival analysis. It's basically a type of logistic regression. It's based on that and it's used to analyze time to an event. So why not use ... that should say used logistic or linear. Actually, there's two here. On the one hand, you can think about ... if you think about variable days to arrest, that is, it's not really a scale variable although its tempting to treat it like one because we had the minimum of one and maximum of about 1000 and all of the units in between there.

So in some way, that could be some people would want to model that in a maximum likelihood model. But that doesn't really work because A, there are no negative values here in regression assumes that the negative value could at
least exist and everybody is censored. And so the data are not truly continuous 'cause everybody in the models is censored at some point. This violates our assumptions of normality which makes it useful then to use a logistical technique on hazard and survival functions. So that's basically why we're doing it. Why survival analysis accounts for all this by determining the likelihood of surviving at each unit step. That's the survival function. So for each unit step, it models whether the person survives or fails and then puts that into the analysis which we see.

Daniel O'Connell: Now for most users in criminal justice, there's only a couple of things in survival analysis what are going to be of interest. The medical folks like to use the life tables which can be produced. But we don't really get into those too much. What we want here are the Kaplin-Meier graphs and Cox Regression coefficient. So we're going to take a look here at the Kaplin-Meier technique and then at the Cox Regression techniques.

Daniel O'Connell: So Kaplin-Meiers are basically a way of running a survival curve. It's sort of a bivariate thing and what it does is it gives us a survival function on ... well you could do one group but generally on two groups is what we're doing. So in SPSS the way you would run this, if you're doing it in the Windows version, you would go to analyze. That'll get you the drop-down menu. We're coming over to survival and then we are selecting out Kaplin-Meier. All right?

Daniel O'Connell: So analyze, survival, Kaplin-Meier and then if we go to the next slide, okay that will give that into this window here, and when that window pops up, we're looking to put in a few things. We want the measure of time, in our case, that's the time to variable that has a value for everyone, the one we created. We want the outcome event, that's our status and that is a NewCrime. When you put this in here, it'll make you click here, the Define Event and it basically just wants to know the unit that defines the event. For us, it's a zero one variable and so we want that one and then the factor or as the comparison variable, that's our treatment variable, it's conditioned in this study. So if we load those three things in, and hit Okay, it'll run.

Daniel O'Connell: I always tell people its really good to get your code down so you can either write your code, or I tell people, even if you're doing the Windows based technique for this stuff, if you click and put it into a ... it'll produce basically this and in a syntax window which is really handy because then you can go back later and see exactly what you did. Also, makes it easier to start tweaking things.

Daniel O'Connell: So anyway whether you do this or click Okay, that will run the model and you will see this. This is a Kaplin-Meier Survival Plot. When we talk about survival analysis, this generally what people are looking for and these are a powerful tool for talking about data. All right? So survival plots then it's a graphic representation of the time to failure. So here's time on the bottom going out and here's the cumulative survival. So in the beginning, everyone has ...
surviving on day one, all right, and then we see people start to drop off and they follow different curves. I'll talk about that in a minute.

Daniel O’Connell: Again, this is utilizing all of the data for everyone. Bivariate analysis presented earlier, if you remember those shrinking samples we have. This one, when you use the Kaplan-Meier, it uses as much data as you have and you look here, the little T's are when somebody is censored. So that means the clock ran out on them. Okay. So you have two things going on. You have the people failing out by getting arrested and then you have the people who are dropping out because they've been time censored but either way, it uses each case for as long as the case is valid. And so one of two things happen. The person fails or is arrested or is censored by time the end of the study period and has not failed.

Daniel O’Connell: So going to the plot, its pretty easy to see. You know we have the two groups here, the blue is the standard probation and green is the treatment and basically what we see here is that the standard probation group starts failing and falling off pretty rapidly. Right? And not only does the treatment group do better, but the shape of the curve was interesting. So it's telling us something that we wouldn't know just from looking at other things. We can an initial drop off, not as fast as the probation but going down then there was a plateau effect, where people stabilized for a while and then that fell apart. All right? So being able to analyze this, somebody can go back to the program or whatever it is, and try to determine what's happening, okay, so at a certain point, people got stable and at a certain point they started dropping out.

Daniel O’Connell: So those are types of things that visual representations can be pretty useful for and one of the things we get out doing these types of survival things.

Daniel O’Connell: Okay. Along with ...

Erin Farley: Dan?

Daniel O’Connell: Yes?

Erin Farley: Yeah, I actually, could you go back, I have a quick question, I just wanted to clarify something. I don't know if anybody else might have this, I just wanted to ... see you're talking about the people, those little crosses represent people who have been censored, so I'm assuming that's people who, let's say they were released very close to the end of the project date but they were included in the project. So they have once that date ended and they were censored, they just have a smaller time frame so they recidivate but they have a small number of days, does that make sense? Is that what those crosses mean?

Daniel O’Connell: Yeah, yeah. So the crosses, you have two things happening and its hard to see in this one. It's a little dirty. Partly because there's a lot of cases in here, so you have the censoring people and then the drop people. So you have the two things happening at once. So following the shape of the curves then yeah, you
are correct. So you've got the censored and then steps. So like, here are the steps and people are stepping and getting censored.

Erin Farley: Okay. Thank you.

Daniel O’Connell: Make any sense?

Erin Farley: Yep. Thank you.

Daniel O’Connell: Yep. And that actually leads in well here because when we get into the tables that come with the output as well, you get the percentage of people that survived, right, so the 39, these are the events, that's the number of people arrested, right? So in the control group, we had 103 people arrested in the standard probation groups. We had 39 people arrested. And because in this one, they're weren't as many people arrested, almost 80% of the sample was censored because they didn't get arrested. So if you don't get arrested, eventually you get censored. Okay? Whereas, in the standard probation group, where there was a higher number of people getting arrested, fewer people were censored, right? So this basically tells you that 80% of the people ran out the clock, in one way, shape or form in the treatment program where there's 48% ran out the clock in the standard probation. So there is a lot of useful information in here.

Daniel O’Connell: Now if we jump down to the means and medians for survival time, we get more information basically looking at the estimates and then the confidence intervals and you get the medians' survival time, and the means survival time, by group so just to take the means here we get like, there's the overall means, 291. The treatment means was higher, and the standard probation means then was even higher. All right? And then we get the 95%, we get the confidence interval bounds around those. So there's a lot that you get out of this table but again this is basically by bivariant stuff. It's bivariate on an outcome, but what it's doing is modeling time in a pretty unique way both graphically, and data wise and because of the estimators that it's using, it's able to take advantage of all the data, which the techniques we look at earlier in the presentation, were not doing. So that's Kaplin-Meier stuff. Okay.

Daniel O’Connell: So that's the first level. Generally when I do this kind of stuff, I do it in exactly the order that presenting it. I start with the data, start playing with dates of getting all of that right, move up through my simple frequencies, crosstabs, and T-Tests. Then I move into my Kaplin-Meier tables and start working with this stuff and then finally, at long last, we get to our Cox Regression models, which in many ways is the meat of survival analysis. This is like if you were doing regression models, you would do the same thing, and the last thing you would do is run your regression models. It's the same here, except we're modeling time.
Cox Regression are an extension of Logistic Regression. As you'll see when we get to the output, it's essentially a Logistic Regression output, and it allows the estimation of time to failure in multivariable models. So what they do is they take ... basically they take this stuff and do it in a multivariable and allows multiple predictors. So we can test the impact of X, whatever variable it is on survival time, while controlling for other variables. And of course, it allows for full model significance tests.

So how do we run a Cox Regression? It's kind of the same as the Kaplin-Meier. As a matter of fact, you start in the same place, analyze, go to survival, gives you the drop-down except and instead of selecting Kaplin-Meier, we're selecting Cox Regressions. You can use Cox with time dependent variables. That does not come under the introduction to survival analysis so we're not going to talk about that today. But it's basically just a different step.

So that, if you go through that, if will bring you to this which is where you set up the SPSS model, much like setting up a Kaplin-Meier. These two, in fact, are the same so we have our time variable. Again, we're using timetoar, we're using NewCrime but because it's not a bivariate, because it's multivariable in this block, we can enter mutivariables and here we’re entering condition, startage, this is a control factor for age at firstdearr which is a criminal justice control variable, which in it is not making much difference in this one.

Anyway, what you can’t see down here, is we have race and gender in the model as well. We’re not using strata here. On the methods, you can do some step wise things. Here you can select for this example. We’re going to answer everything and we’re going to answer it all in one block. If you want to get fancy, you could often, you'll enter things one at a time, so we could do block one, enter a condition, block two, if you would click here and then build your model up over blocks. And it will test the significance of those blocks as you go forward. But in the interest of keeping things simple, we're just going to do a full answer model here. And again, if you click paste or write the code yourself, you'll get this. If you click okay it will just run the model.

And so the first thing we get are the full model test results. We want to know if our model even fits the data. Generally they do. But we want to run this test. So the way this works, it’s a -2 log likelihood model. As you see in some other methods, we use these types of things. There's no real interpretation by itself of a -2 log likelihood statistic except that's smaller is better.

The key is you could look at the difference between two model estimates, so model one and model two, you can look at the change in the log likelihood which follows a pi squared distribution. So we have our model. What it does, when you run in SPSS, or any other program, it runs the model with no predictors essentially. It gives log likelihood for that. And then the model with the ... all of the predictors and it gives you a second log likelihood for that, right? And essentially, the difference between those two is pi squared statistic. So the
difference between the two, you have the overall, the change from the pre ... this is actually the one it's modeling, to the change from the previous step and that's 55 ... it's a typo, it should be 59.937 and of course in a pi squared distribution that's significant, no problem. All right.

Daniel O’Connell: So we know our model is significant. We know it's a valid model and therefore we can go to then next step and look at the what it's showing us. And these are the individual variable tables. These are saying, if you look ... if you just jump down here and look at this thing, if you're used to logistic regressions, this is going to be a very, very familiar table because that's what it is, except the exponential beta is the probability of experiencing an event throughout the observation period. This is the Hazard ratio as opposed to an Odds ratio. So its exponential beta, referred to as Hazard ratio, right?

Daniel O’Connell: So what that means? An exponential beta here, a Hazard ratio above one means that for a one unit increase in X there is a corresponding increase in the probability of Y occurring. All right? So for us, I picked a positive one even though it's not significant from our data. Being male increases the probability of being arrested by 1.48 times. That is this. All right. So that's a positive exponentiated beta. Okay?

Daniel O’Connell: For our condition variable, which is really the variable and that's holding everything else constant here. In this one, the only really thing that jumped out was the program variable condition, it was negative. For our condition variable, which was .405, so less than one so it's negative. We would say that the odds of someone in the program, not surviving, not getting arrested, are .405 what they are for those not in the program, controlling for variables. I have always found this language incredibly confusing, and in the literature and just in general. People disagree on how to interpret, what is the best interpretation of these negative exponentiated betas.

Daniel O’Connell: And so I do what I always do, I take the inverse. If you want to flip a negative logistic co-efficiency, just take one divided by your co-efficient and that may gives you a positive. In this case, 2.4. So then the language becomes, the odds of someone in the program SURVIVING, not getting arrested were 2.4 times the odds of someone not in the program, controlling for other variables. So that is essentially it.

Daniel O’Connell: So we move from basic very simple things, modeling time, getting our dates straight up set, getting account of the length of time it takes for something to happen, the moving ... I'm going to jump through the slides here real quick and probably confuse you all. Moving to these Kaplin-Meier models, where we have our survival functions, where it starts to get fun. There's are kind of our bivariate models of time. Then moving to the Cox Regression models and ending here with our output from the Cox Regression. All right.
So somebody asks and somebody does ask, why didn't we just use a logistic regression on the zero one whether people failed or not? Because we would've had a lot missing data for one thing. The main reason then is it's survival analysis with Cox Regression, it utilizes all of the data by modeling that censored end point, every piece of data in the ... that have collected gets utilized right? And our data the end point was the end of the study but even the best administrative data as an end point that's today, if nothing else. If we're looking at people that we're doing a study and today I April 26th, 2018, those people are all censored as of today. Now, we don't know whether they're going recidivate tomorrow, all we can say is that they haven't recidivated. And that's all time dependent data are essential censored. And survival analysis accounts for this. It allows us to deal with this and to make a more robust and a more confident statement about what's going on with our data.

So that's basically it. So. We're kind of at the end. People are often asked to provide reports on issues that involve time dependent variables. People want to know how long it takes for something to happen and what predicts how long it takes for something to happen. That's where these things come in. Survival techniques provide relatively simple ways of doing ... answering these questions. I think these models are fairly straightforward. I think that ... you know the trick is in the data on this stuff. As I said in the beginning, date variables are very, very particular. Every time I do these types of analysis, the vast majority of the time we spend is on getting the date variables correct.

Once you've got the data and the analyses are fair straightforward. Once you've got that, I echo this every time I do a presentation or in a class, I say start simple, work your way up. Don't jump right into the survivals. And the other one I always conclude with, is remember you're not alone. Do try this at home. If you got some data, pull up SPSS or Stata and just start playing around with some stuff. Your colleagues are your best resource. Go down the hall. Talk to somebody at your agency. Use your resources. We live in an age of researcher/practitioner relationships. If your at a SAC and you've got a local university, call those folks up. If your at a university and you've got a SAC, call your SAC 'cause they've got the data, and often they're for new way to analyze it.

More and more, you can get questions answered on Google. There are forums everywhere where you can go ask statistical questions and get answers to them. There's tutorials. There's How's To. And reach out to people via email. And that's about it.

Looks like we have a question from Ryan.

Erin Farley:

Yeah.
Survival Analysis: An Introduction
April 26, 2018

Daniel O’Connell: I'm interested to know a little bit more about the consideration of censored data within these models 'cause the regression weight cases by a length of time into the study or does it just not omit cases if they are censored before ...

Daniel O’Connell: It doesn't weight them. It gets a little funky in this. Basically, it's not omitting them for the censor. So its modeling two things at once, it's modeling either time to failure or time to censor. And the underneath of that gets a little technical. And frankly, I don't even always understand that. I've got to look it up. That's a bad answer to that question. But what it is doing is dealing with that right-hand censoring in the algorithm that is running. And yeah, to get into that, is probably a little bit more advanced than today. That's a great question.

Erin Farley: I have a question, oops, Dan, another person, let me ... that you might not see. Let me see. I accidentally just got rid of it. Okay. Here we go. Okay so here's another question for you. Do you subtract one from date arithmetic so that those released and arranged next day are one and those release and arrested same day are zero. Let me know if you want me to repeat that.

Daniel O’Connell: No, I got it. I got it. I got it. In this one, we didn't have to 'cause we didn't have that, but yes you could subtract one or add one conversely, so that you don't have any zeros in there. You probably don't want any zeros in your time. You want everybody to be a positive. So adding one or subtracting one depending, which would be valid. I've never actually had that happen. But yeah. So yeah if you got a zero in there because zeros in these things, what would it do if you zeroed out? It would kind of fun. Yeah so when [crosstalk 00:48:18]. Ah, go ahead.

Erin Farley: Oh no. I'm sorry. I keep explaining, but I just want to being your attention, there's another question when you're done that you might be able to see as well. Sorry.

Daniel O’Connell: Oh. Okay. Well on the other one, I never actually run into a case where I had a zero when I did the data arithmetic. So I'm not quite sure what ... my hunch is you ... you would either add one to the days once you got it, if you add a zero to avoid that problem. Or you could add one to the arrest date to avoid that. I think that's right, yeah.

Daniel O’Connell: Are there any applications for modeling recidivism with a survival modeling seeing is how the assumption here is not that everyone will eventually will recidivate as in a traditional clinical use of survival. Not assuming everyone will die, does this affect our interpretation of ... have you ever had to justify this?

Daniel O’Connell: Well I'm never done it on a straight recidivism study. I've always done them on survival studies where the study ended. So we had an end date and people were getting ... were ending there. So it wasn't necessarily a death, but for everybody in the model, the thing ended. I guess you could consider a recidivism study the same way, 'cause if I'm doing a three year recidivism study, essentially that
three year window is my right-hand censor and it's not that everybody dies at the end of year three, but the clock runs out on them. I've never ... so no I've never had to explain that. But I assume that right-hand censor on a recidivism study would be the same as it is, so we make the assumption, not that everybody dies, but that the clock runs out of them so everybody gets censored.

Daniel O'Connell: I'm not sure if that made sense, but if I interpreted your question right.

Erin Farley: Great. I'm going ... let's wait a couple minutes to see if anybody ... Jason has got things. As the-

Daniel O'Connell: Jason or anybody else, if you want to email me we can talk more about this actually.

Erin Farley: We're going to launch the poll while everybody's still on. Jason, is you could launch the poll, that would be great and then we could see if anybody has any additional questions. Great.

Daniel O'Connell: I take it I'm not supposed to fill out the poll.

Erin Farley: No. Sorry.

Daniel O'Connell: No, no. That's okay. I just kind of funny that it that didn't even popped up.

Erin Farley: It would be like, oh she's wonderful.

Daniel O'Connell: Well actually I was thinking I confused myself.

Erin Farley: So here's a question, I know that you usually use SPSS but are you aware of like are there pros and cons to using it and the Kaplan-Meier or Cox Regression in Stata or do you know of anybody else in the field that has preferences that's not just ... I know that some people like Stata. Some people like SPSS but are there any differences and how it's run or analyzed in the system?

Daniel O'Connell: I don't know that there are differences. But I know like the guy down the hall from me like literally three doors down the hall from where I work, runs these things in Stata. And he and I go back and forth with output all the time. And output is essentially the same. Essentially, I mean the key components are there, each one gives you some different things that I haven't looked at his Stata ones 'cause he'll come down the hall and show me some plots and some tables and I can interpret those. My hunch is the algorithms are pretty much the same. Then what would ... and with all of these things, people work in the program, they're most familiar with generally. But I don't think it really matters which one you use.

Erin Farley: Okay. Okay thank you. Great, so while we are just wrapping up, I just want to again, thank Dan for hosting and leading this webinar today. If you have any
questions, you can reach out to him via email. And I would like everyone, to thank everyone for attending today. We do have two upcoming webinars as I had mentioned earlier. Just want to advertise those again as we're wrapping up. We have one on May 3rd which is next week. So that one registration is available and open and then there is one scheduled for June 14th and that on should be available soon. So keep your eyes open.

Erin Farley: And I guess if no additional questions, I want to thank everyone for joining us today. And have a good afternoon.

Daniel O’Connell: Thanks.

Erin Farley: Great. Thank you.