ADS-506: Applied Time Series Analysis

Lab 2.2: Differencing and Regression Models

Hi everyone. Welcome back. So in this lab we will be covering the easiest way to remove trend or seasonality and we will also be covering linear regression with time series. So I'm just going to paste a couple lines here so you guys not have to watch me type. So all I'm doing is I am going to load my library. I just got through packages. So we're going to need for this exercise. And then we will then, after we load those libraries, we will load in our data set. So I have some ER arrivals and then going to convert it to a time series object. I'm not specifying the frequency here. If we just look at this ER arrivals ... oops.

So we can see if we have hours. And so I can specify frequency. I can either do 24 for 24 hours in a day. I can also do seven days. This data set only has I think 17 days. So if I do seven days, we have two full seasons. I would be hesitant to do that. I usually like to have a few more seasons in my model. So the minimum would be two, but I like to have a little bit more. 24 hours, we could do that. If I don't specify, then there's just no season. And if we just view this, so I'll enter the keyboard, we just see that there's 408 observations and there is just a frequency of one just meaning that I did not specify a season. So as far as the computer's concerned, there is no season. All right. Let's remove seasonality. So there's a few ways to do it. If we just view our data right now, so we definitely see this. There is seasonality here. We can just see the spikes every 24 hours. So what we can do is we can do what's called a different scene. So we can take a first order difference.

If I could just type it. And all that is going to do is take your observation at time T and subtract your observation at time T minus one or actually time T minus whatever we specify is our difference. So we could do lag one. So it just, like I said before, is time T minus the value at time T minus one. We can also remove season. So it'd be time T minus at time T minus whatever your season was, in this case, 24. So our value at time T minus the value at time T minus 24. And then the difference of that is our new value. That is what we will be using as a new time series object. So if I just run the disc function and then I can say lag equals 24.

So if I just run this without saving it, you guys can see what it's doing is we actually have, so zero ... let me lower this a bit so you can see. So it was the time this minus whatever the time was 24 hours ago, which looks like that number, and then we have a nine minus of four. Sorry. It's a four minus of nine. That's how we get negative five and the three minus of six. That's how we get negative six. And so forth. So the other thing if you notice, we are

now starting at observation 25. And that's because the first 24 observations were removed. And so now all we have is just a difference and got it saved.

We can plot the difference. And it looks like that. Okay. And it's close to what we want. We definitely want our mean to be around zero and it's close to that. You can also sometimes, you'll need to, is take a second order difference. So trend, if you want to detrend something, that's usually just lag one. And then if it's seasonal, then it's lag whatever that season is. Sometimes your data will have both season and trend. In that case, you'll need to take a difference. Right here this is a seasonal difference. And then I can also take an additional difference.

And so this would be a second order difference. You should never have to go past that, like third or fourth. It's either first or second order. And like I said, if we're detrending it, then it's lag one. If we're removing the season, then it's lag whatever that season is. All right. So let's make our training and test set.

All right. So I'm going to make a training set based off my difference data and difference is one. So I know it's going to end at observation 384 because there's 408. That's the end of it. What I'm going to do is I'm just going to make the first part of that is my training and then at the very last day I'm going to use as my test set or validation.

So that will start on 385. And the reason I'm starting at 385 is because 385 plus 24, we're going to get to 408. Okay. So once those are done, let me just run these. Oops. All right. Now let run these. So my training set and testing set. All right. So we can now make our model.

And I'm going to use the ETS function. Okay. So it's exponentially and trend and season is what the ETS is. And so that's ETS. All right. And then I will say my training set. And so I have to give it a model. Now you'll notice here in the tool tip, the default model is three Z's. Z, Z, Z. Basically what that means if I left it alone is R will figure out what that model is and it will give me the best model it thinks best represents the data and we can force it. I can just say, "You know what? It's going to be A and N." And so what I'm saying is that the A is going to be for additive, the second N means no trend, and the third N is no season.

And since we took a difference, there wouldn't be a season. That's why I have that A and N model. All right. And then I can specify my alpha and this will be the weight that this exponential smoothie will use. And if you're wondering how we come up with the alpha, it's really just coming up with a ballpark. So actually I could just use it to eight. The closer you are to one, the more that all your lag values are weighted equally, then we could say the

most recent lag is weighted more. And so it exponentially drops off after that.

All right. And this is just no season and no trend. That's why I chose the A and N. We have the prediction. So just forecast.

And I'm going to forecast the next 24. All right. So let's see what that looks like. So here's our model. And so it gives us the AIC and BIC values. If we had other models, so if I wanted to, just for giggles.

All right. So we saw the eight. So let's see the four. So the eight, the AIC is 3272, the 0.4 is 3179. So the 0.4 is actually a better model because we want the lower AIC score. And the 0.1 is a 3135. So the 0.1 is actually a better model. Now you can get to a point where there's not much of a difference. So going from 3135 to 3128, it's not much of a difference, but going from 3179 to 3035, that's a pretty good difference. So I'm actually going to change this back. Let me remove these. And I can run these again. And so these are going to be our prediction. So for exponential smoothing or weighted average, what this is doing is it's looking at that last value and it's pretty much just keeping it because it has no new information.

So it's just taking whatever the last weighted value was and just keeping it consistent. So if we were to apply that, we would just see a straight line going across for the next 24 observations. Actually we could plot that. Let's pull up the train.

All right. So what I'm going to do is I'm going to layer the test, which will be our actual. And then we're going to auto layer our prediction and that will be our prediction.

It's helpful if you type it in correctly. All right. Cool. Ooh. Can't see that. Hold on. Let me make adjustment. All right. Actually I'm going to do one more so we can zoom in.

Let's go 375 to 415. All right. So all I did was zoom in. I did not change the plot at all. So here we go. Here's the training, which is the blue line and then here is the actual, which is this red line and then this green line is what we predicted. Now I included the confidence interval. So you can see this green line is what we predict it will be, which is basically just the weighted average as a previous value, all the way across. Now it's within our 95% confidence interval. We actually have 24 observations. So there's like two points out. Those aren't really statistically significant because we would expect for every 20 observation about one to be out. So if we saw three then something would be off.

So we're good with that. And so that's just using the exponential smoothing

and weighted averaging to make our predictions. Okay. So this plot is just showing us the difference value. And again, if we look at our predictions and we're looking at that, and again, our data set is ER visits. I don't know what 0.26 is. Like how does that relate to an actual person coming into the ER? So what we could do is we could just convert it back to the original scale.

Let me just show you how. I always found it a little bit annoying when the books would show you. Okay. Well here's how you do the difference. Here's how you make your predictions and you're good to go, but then you never actually go back to the original scale. All right.

It's a really long title. All right. I'm going to make it a time series object because I'm going to plot it with our original. And the original data is a time series. So in order to plot them, I need them both to be time series objects. All right. So really all we're doing is we're going to take the last value that we know, which is 384. That's the original value. Oops. Let me score brackets. Original value is 384. And then we will add the cumulative sum of our prediction. Now remember, our prediction has all these values. So we actually had to specify and with the dollar sign, we can specify what we want. So we specifically want the mean, which is what this first column is.

So that and then, because we're dealing with timeshares, we have to tell it when to start. We are not starting that observation one because if I start at observation one, when I plot it, it'll be at the beginning of the graph and I actually want it at the end. I want to start right after my previous value. So this previous value is where my training ended. So it ended at 384. So I'm going back to my original data set, my original arrivals, and I'm going to observation 84 and what I'm going to start doing is just adding the cumulative sum. So it would be whatever the value was at 384 plus 0.26. And then at 386 it'll be whatever it was at 384 plus this 0.26, plus this 0.26, and just keep adding over and over and over again.

All right. And that's what that looks like. So these are the actual prediction after, I guess, un-difference the data. So let's look at that.

All right. Let me zoom in again. Actually I'll just copy this. All right. So there we go. And that's actually what we would expect. Because we are smoothing out our data, we would just expect a nice line. And so this one has a line going up. So it wasn't too bad. Probably in the early morning it was closed. Mid morning to afternoon, pretty off. So that is taking the difference. Like I said, first difference, usually when we want to take away trend, it's just the current value of time T minus the value at time T minus one. If we want to remove season then we would take the difference of YT, the observation at YT, minus observation at YT minus K. K being the season.

All right. Cool. Let's go over regression. So we can do regression with a time series. So I'm going to say new arrivals. All right. So all I'm going to do is I'm going to take ... actually. Give me one second. Sorry. I already have a rebels. That's not original. Yeah. All right. Cool. So let's make a new training set, test set. Action. So I'm actually going to go back and give it a frequency. I'm going to say, "Okay. This does have seasonality. The season is 24. 24 hours in the day." So we have 17 days and each day is its own season. All right.

All right. So I'm going to make my linear model training set, which is just you got to probably forward this. You should know it. All right. So now, because I have a season and I specify when it ends, I'm not just putting in a number, I have to tell it, "All right. So the 16th day and the 24 hour."

So task will be reveled. So now the tasking will be on the 17th day, first hour, hour one. So now these exist. All right. And so to do a linear model now, before when you had a data set and it was multi-variate, you would just do LM. If you were using R, you would do LM, then your outcome variable, and then predictor one, predictor two, predictor three, so on, and then your data set. That's original. That's what we're doing. So with time series, we're going to do pretty much the same thing except we're going to add something. So let's make our model, arrival. We are actually going to say TS for time series, linear model, and we are going to get arrivals.

Now I don't have any predictive variables at all. So I'll come back to that, but my data is just arrivals. So what I do have with the time series linear model, we can specify trend. Basically we are telling the computer to look for a trend or we know trend is there so find it. And we can say season. So they're default. So we can put that in there and we can actually have it look for it. I'm actually going to remove trend just because these arrivals don't have trend, but they do have season.

And we can just run that and what it does is just like a regular linear regression model, it gives us our Y intercept and then coefficients for each predictor variable. In this case, the season is going to be our predictor variable and it's basically telling us the relationship of this hour to our arrival. So starting that midnight at zero, we can expect 3.8 patients according to this model. And then the next one, even though it says season two, it actually means 1:00 AM. So 1:00 AM is actually going to go down by 0.17 and 2:00 AM, 3:00 AM, 4:00 AM. So here around 9:00 AM, from whatever number we had at midnight, we can expect five more people. And again, at 11:00 AM, whatever we had at midnight, we can expect six more people and so on.

Those are ponents. All right. So I just ran the summary function. I like using

the summary better because it puts it into a nice, neat columns for you. It also gives you your residual standard error. Our residual standard error is 2.9 and then it gives us our R squared. So by having nothing else, we can say that just knowing seasons, that can explain 54% of the variation in our data. Obviously there's other variations or other outside forces that are affecting our data, but according to this model, we only have seasons and it can explain 54%. All right. So let's make some predictions. All right. So we're going to use a forecast function.

And we want to predict 24. Oh. So here is our prediction for the next 24 hours and we can plot this to see how it looks.

So we're going to pull. So it's going to be our training set.

So now our prediction.

[inaudible 00:33:20] All right. There we go.

So let me zoom in.

Okay. So let's zoom in.

I want to go from day 16 to day 18.

Our model. Oh. My fault. Sorry about that. Okay. So my time series linear model, I should have used my training set. Instead I used the actual data. So let me rerun this. All right. I was wondering why my was off.

Okay. So that works.

It's still giving me an error. So let me just ... all right. We'll run that. I just went through my original data. Okay. I got to look into this. I'm not exactly sure why, but it's off by a day. All right. So let me just reset everything. You know what? Since it's off by a day, then we'll go put one day forward. All right. So for some reason my prediction is going off of day 18 instead of day 17. So, anyway. This is what it would predict.

Well, I was going to say we could plot it, but that's the idea. We can look at what our predicted values are and we can see that, assuming this was day 17, we can see midnight, 1:00 AM, so forth. And we can see three, four, five, six, seven, eight, nine, 10. So around 11, 12, and 1:00 PM we can affect like 11 patients. And so if we looked at that, it could be closed and we see the dip. So we see the increase and then a dip and an increase, a dip, an increase. So that's what we see here. And it's just in the actual, it was definitely more pronounced. In our model, not so much, but it's still there.

All right. I'm going to show you another linear regression model. I'm just going to import this fresh. And then what I want is I'm going to make some dummy variables. All right. So let me do this. So dummy variables are going to be variables that we impute, variables that we can gain from the data itself. Since we're dealing with dates, we have all kinds of variables we can create. Obviously days of the week, weekday, weekend, we can create season like actual seasons. So winter, spring, summer, fall, evening, day. We can create all kinds of variables just on knowing the date and time. I'm going to create one real quick.

And just I'm going to create day of the week. So that's going to be date. You got to format the date first. So format date with the percent A. That's going to give us the abbreviated day. So like SUN for Sunday. Because I'm creating a factor, that's basically I'm creating a categorical variable.

Sunday, Monday, Tuesday. And I'm specifying what order. If I don't specify what order and I create a categorical variable with the dates abbreviated, then it'll put it in alphabetical order and I don't want that. So I have to tell it what order to go in.

And I don't need my date anymore. And let's already convert this to a time series serious.

All right. So I have a new data set. Arrivals done. And we can see here, this is a multi-variate time series. And all that means is I have more than one variable. So arrival, that's still going to be my outcome variable. I now have a predictor variable alongside with my time itself. All right. So now same thing. Let's make our training set.

And this will end, again, same day, 16th day, 24 hour. This will start the next day, 17 one. That's that 17. So it ends on 16. All right. So arrivals, some linear model, BTS LM. So this time I will put arrivals because that is what we're trying to predict. I also have a day and I believe it's capitalized. All right. Just look at my code since I wrote it. Right there. Day and then I'm also going to say it has season. My training set.

All right. Let's run that. And so, what was the one before that? So before our model is 5394. This model is 5342. So we actually explain less. So we pretty much just gave it some trash. Interesting. But for the purpose of the demonstration, we will continue. All right. So now if we wanted to make our predictions, we got to tweak this because before we can make a prediction, I could just use the forecast and then put in my object and say, "Okay. Now give me the next 24." But now it's a little bit different because I actually have a predictor variable. All right. We see that the date is three. So it's going to be the third day. All right. I'm going to make a data frame then call it next

day.

And so because I have to tell it with my prediction, I had to provide the same variables. So if I had day and then, I don't know, month, so on, then I would have to provide those for my prediction because those are factored into our model. So it needs to know what those values are. So we have our data frame. Day is equal to, we got to repeat three 24 times. All right. So let's do this.

So we're going to do forecast, same thing. And we are going to use our model, the dum model. Seriously. Dummy model. Okay. Now this can be a little bit different. We're going to say, "This has new data." And our new data is whatever we set for next day.

Okay. So it took the hour of the day and then it said, "Okay. Well it's going to be the third day. So that's going to factor in somewhat." And so if we want to know what the third day the coefficient was, so three times negative 0.62. And we're going to add that to the value that we have at midnight. I'm sorry. No. That was the hour. The day is negative 0.08. So it's going to be negative 0.08 times three. So really not much of effect. And then we're going to add that. And so this is what our prediction is. I'm going to try to plot this one. Make no promises, but kind of fun to watch your teacher type and fail. All right.

The auto layer.

You don't have to give it a series name. I just do because it's easier to see.

And our prediction.

All right. So let me zoom in. Clean this up a bit. That area you're seeing, because I do have a multi-variate time series I didn't specify which one I want plotted to plotting them all, which is the hour and emphasize the day of the week, but right now I don't care.

And so let's go for ...

Okay. I'm just going to zoom in some more. So this is what our actual is, this red line. And then we can see on the test day, our model's actually pretty close. It's this is blue line and the red is still within our confidence interval. So it looks like around one. We thought there'd be a little bit of dip, but in actuality there was an increase, but it's still the same overall pattern. So this is how we would do time series linear regression model. And just like other linear regression models, we would look at the summary, we could look at our residual standard error. And again, when we introduce new variables, we

would want to look at the residual standard error to see which model has the lowest error because that's the one that we're going to want to pick. Okay.

A lot of people look at the R squared because that explains your variation, but fun fact, the more variables you add, the higher your R squared is going to get anyways. Even if you use the adjusted R square, which takes into account the number of variables you have, it says an increase. So even if your variables have nothing to do with your outcome variable. So it's safer to go with your standard error. You want them all with the lowest error when you're comparing models. Cool. So that is the end of this lab and we'll see you next time.