Individual assignment SBD

Michal Kubina

ADS, 2021-2022

Contents

0. Prepare

▶ Load the R-packages you will use.

```
library(fpp3)
library(tseries)
library(readr)
library(expsmooth)
```

▶ Include R-code you used to load (and prepare) the data.

```
data <- read_csv("pat_205.csv")
```

```
## Rows: 131 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (10): mania, depression, actipoints_median, actipoints_min, actipoints_...
## date  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1. General

▶ To be able to use fpp3, the data have to be a tsibble object. If they aren't already, transform them. Describe the structure of this object.

```
data <- as_tsibble(data, index=date)
```

The tsibble object adds an indexing parameter to time series data. Basically, it creates and preserves the index of the data frame. It makes working with specific functions better and it is very similar or even an alternative of the pandas object in Python3 where the index is also preserved.

1.1. Describe your data

My data come from a company called Mindpax.me. The data are from a patient diagnosed with F31 disease which is bipolar disorder. The patient wore a wristband for 131 weeks which captured actigraphic data. The actigraphy data are essentially a time series with a fixed sample rate where each sample corresponds to a particular value. Specifically, the data collection was achieved by a special wristband called MindG (https://docplayer.net/59910788-A-comparison-between-the-mindg-wristband-and-the-motionwatch-wristband-for-monitoring-physical-activity-and-sleep.html) with an accelerometer included and provided, as already mentioned, by MindPax Ltd. This device collects signals from three axes (x, y, z), and then the magnitude of the acceleration vector is computed as a squared root of the sum of the signals from the x, y, and z axes squared. The acceleration vector with the highest magnitude is chosen every 30 seconds.

These data were already aggregated and processed into the weeks from the company by using logistic regression to detect sleeps and decision tree to assess the number of activity points from actigraphy data in a day. Thus, for each week there is a median of sleep (median value from 7 days where each each day there is a sleep duration in hours per day), median of activity points, max and min value of sleep in a week, max and min value of activity points in a week and absolute mean difference of sleep and activity points in a week. To be specific, activity points range from 0 to 3000, where 3000 is the highest possible activity. Moreover, the patients fill a special questionnaire that assesses the last week of a patient from depression and manic point of view. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8386400/). The questionnaire is filled every Monday and the data come from a time period the Monday before to Sunday.

To sum up, my data set consists of the self-reported weekly values of depression level of the patient, mania level values, median values for sleep, and activity points for the corresponding week. All other variables I will leave out and not consider Moreover, it was essential to get the absolute mean difference which captures the total change of sleep and activity during the week. For this, I used the absolute mean difference. (https://en.wikipedia.org/wiki/Mean_absolute_difference)

- depression - range [0 (ok state), 16 (high depression)]
- mania - range[0 (ok state), 16 (high depression)]
- actipoints_median - median of day activity points in a week
- actipoints_max - max of day activity points in a week
- actipoints_min - min of day activity points in a week
- actipoints_diff_mean - absolute mean difference of activity points
- sleep_median - median of day sleep in a week (hours)
- sleep_max - max of day sleep in a week (hours)
- sleep_min - min of day sleep in a week (hours)
- sleeps_diff_mean - absolute mean difference of sleep (hours)

There are no missing values.

```
str(data)
```

```
## tbl_ts [131 x 11] (S3: tbl_ts/tbl_df/tbl/data.frame)
##  $ date                : Date[1:131], format: "2018-10-14" "2018-10-21" ...
##  $ mania               : num [1:131] 2 4 1 3 9 4 7 3 0 0 ...
##  $ depression          : num [1:131] 9 7 11 15 0 5 5 11 6 16 ...
##  $ actipoints_median   : num [1:131] 1225 945 920 1010 995 ...
##  $ actipoints_min      : num [1:131] 1075 860 860 785 995 ...
##  $ actipoints_max      : num [1:131] 1375 1140 1090 1050 995 ...
##  $ actipoints_diff_mean: num [1:131] 300 110 113 117 140 ...
##  $ sleep_median        : num [1:131] 5 5.94 6.47 6.33 6.15 ...
##  $ sleep_min           : num [1:131] 3.23 5.11 4.93 4.83 5.15 ...
##  $ sleep_max           : num [1:131] 6.78 7.62 7.45 6.91 7.16 ...
##  $ sleeps_diff_mean    : num [1:131] 3.542 0.926 1.11 0.944 2.008 ...
##  - attr(*, "key")= tibble [1 x 1] (S3: tbl_df/tbl/data.frame)
##   ..$ .rows: list<int> [1:1]
##   .. ..$ : int [1:131] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..@ ptype: int(0)
##  - attr(*, "index")= chr "date"
##   ..- attr(*, "ordered")= logi TRUE
##  - attr(*, "index2")= chr "date"
##  - attr(*, "interval")= interval [1:1] 7D
##   ..@ .regular: logi TRUE
```

```
summary(data)
```

```
##       date                mania          depression     actipoints_median
##  Min.   :2018-10-14   Min.   : 0.000   Min.   : 0.000   Min.   : 730.0
##  1st Qu.:2019-05-29   1st Qu.: 0.000   1st Qu.: 2.000   1st Qu.: 962.5
##  Median :2020-01-12   Median : 2.000   Median : 5.000   Median :1020.0
##  Mean   :2020-01-12   Mean   : 3.126   Mean   : 5.779   Mean   :1028.3
##  3rd Qu.:2020-08-26   3rd Qu.: 5.750   3rd Qu.: 9.000   3rd Qu.:1077.5
##  Max.   :2021-04-11   Max.   :12.000   Max.   :16.000   Max.   :1620.0
##  actipoints_min   actipoints_max actipoints_diff_mean  sleep_median
##  Min.   : 270.0   Min.   : 970   Min.   : 73.33       Min.   :2.942
##  1st Qu.: 705.0   1st Qu.:1148   1st Qu.:140.36       1st Qu.:5.779
##  Median : 820.0   Median :1225   Median :173.33       Median :6.492
##  Mean   : 800.6   Mean   :1247   Mean   :192.08       Mean   :6.442
##  3rd Qu.: 892.5   3rd Qu.:1330   3rd Qu.:229.05       3rd Qu.:7.221
##  Max.   :1620.0   Max.   :1625   Max.   :503.33       Max.   :8.892
##    sleep_min        sleep_max       sleeps_diff_mean
##  Min.   :0.5667   Min.   : 5.658   Min.   :0.6516
##  1st Qu.:3.0583   1st Qu.: 7.450   1st Qu.:1.2794
##  Median :4.3167   Median : 8.217   Median :1.6597
##  Mean   :4.0464   Mean   : 8.382   Mean   :1.8293
##  3rd Qu.:5.0708   3rd Qu.: 8.933   3rd Qu.:2.1762
##  Max.   :8.0500   Max.   :20.958   Max.   :7.6562
```

Start with answering the following questions:

▶ What is your outcome variable; how was it measured (how many times, how frequently, etc.)?

I would like to forecast and see the causality of depression levels of the patient. Thus, my outcome variable will be the self-reported depression level of the patient which was taken 131 times every week on Monday. This value can range from 0 to 16 where zero corresponds to value with depression symptoms and vice versa (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8386400/).

▶ What are the predictor variable(s) you will consider? Why would this make sense as a predictor?

My predictor variables will be the median of sleep duration in a week and a median of activity points in a week. Considering the literature (https://onlinelibrary.wiley.com/doi/full/10.1111/pcn.12688), circadian rhythms are very important in the life of people with bipolar disease. Thus, the duration of sleep and activity can be quite significant for the depression level of patients. Moreover, even though it cannot be considered a cause, a good predictor might be mania level of a patient. Since bipolar disorder fluctuates between mania, depression, and remission.

▶ What are the cause(s) you will consider? Why would this make sense as a cause?

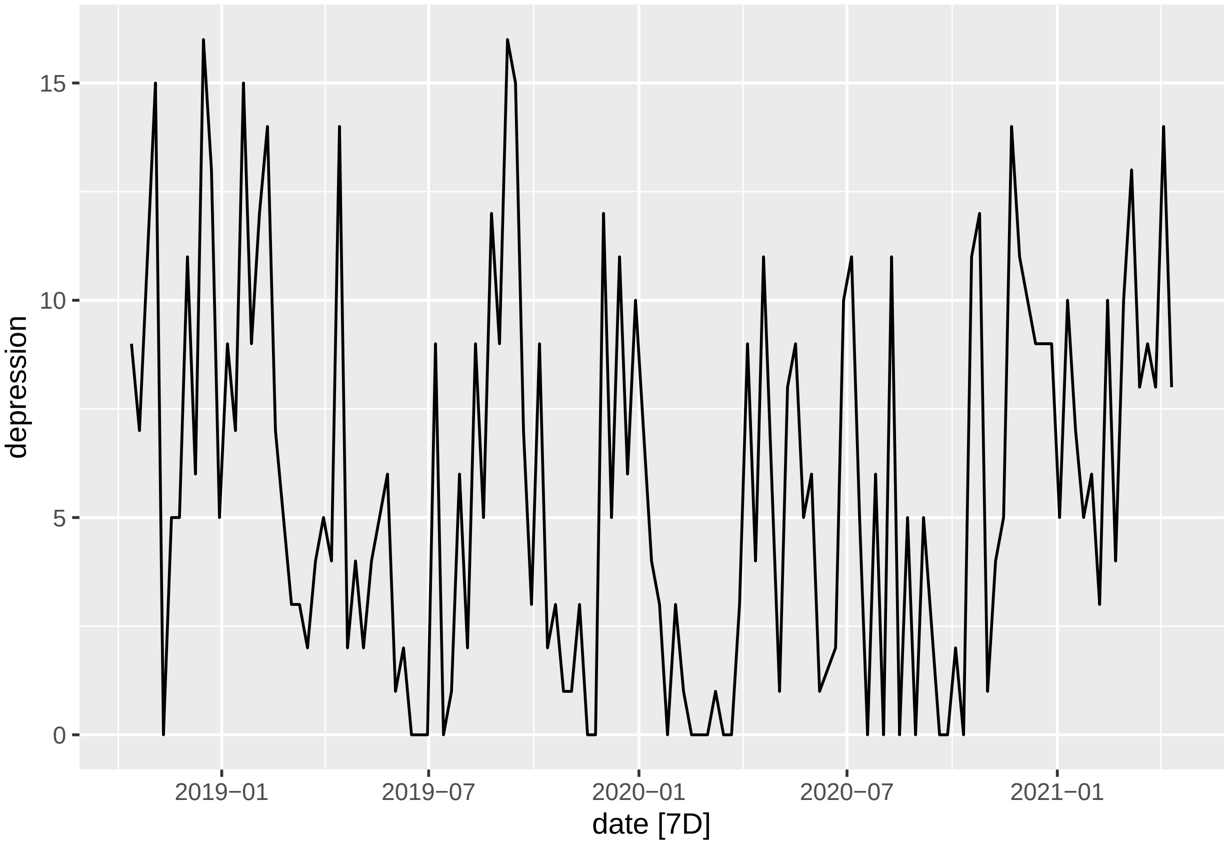
I believe that both variables taking into account sleep and activity points per day can be causes to the depression levels of the patient. I believe that depression values with the values from mania assert might correlate but it should not be a cause.

1.2. Visualize your data

▶ Create a sequence plot of the data with the function autoplot(). Interpret the results.

```
autoplot(data[, c(1,3)])
```

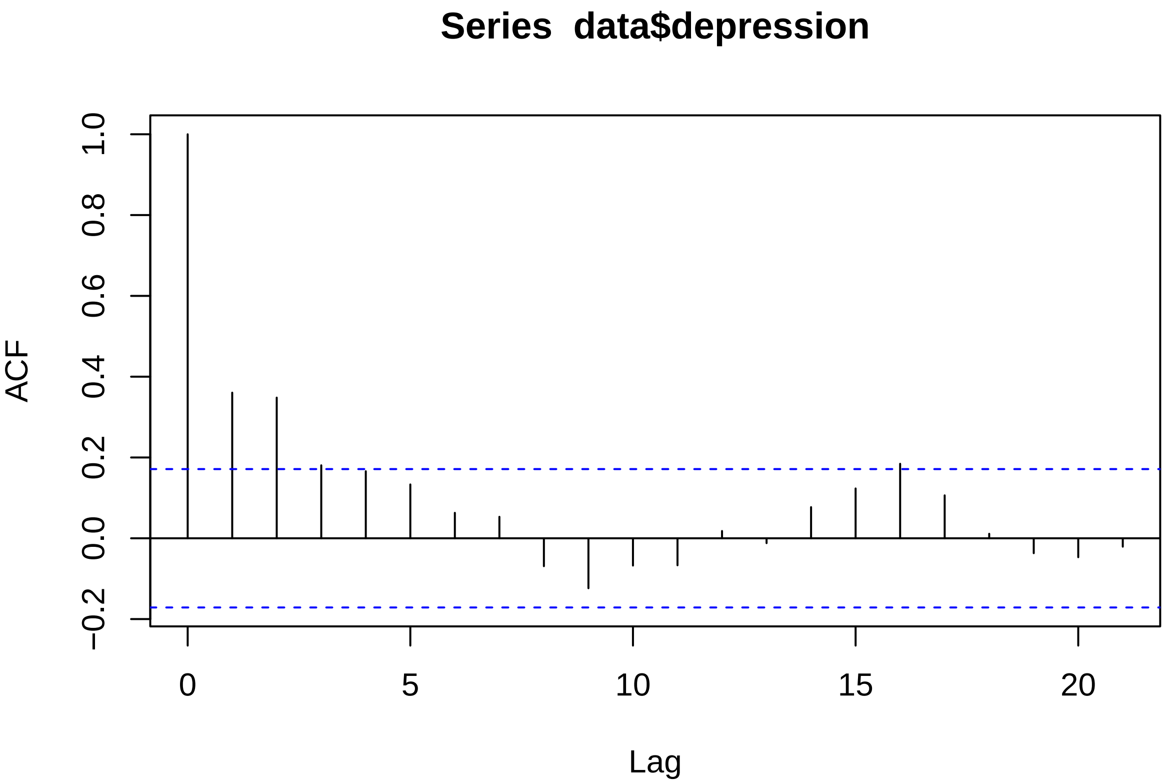
```
## Plot variable not specified, automatically selected `.vars = depression`
```



To remind, the depression ranges from 0 to 16 where 0 is the value with the lowest self-reported depression symptoms. I would say that the plot seems to be quite stationary not following any trend. However, we can see that if the depression values are going higher or lower, it is rather a continuous change. Moreover, more depressive values seem to be in the winter season where there is snow and bad weather in the Czech Republic (the subject/patient comes from the Czech Republic).

▶ Plot the autocorrelation function with the function acf(). Interpret the results.

```
acf(data$depression)
```



Within the first two lags, the correlation seems to be quite significant. This means that consecutive values have an effect on each other. However, there is no significant trend or unit root process. Thus, I would say the series seems to be stationary with very small non-significant correlations after lag 3.

▶ Based on (basic) content knowledge about the variable, and these visualizations, is there reason to assume the data are non-stationary and/or that there is a seasonal component?

To sum up, there does not seem to be a significant seasonal component looking at the plots above. However, one could argue that depression values are higher during the start and end of the year or in general that the depression comes during autumn and winter. This might be because of the weather and mood of people as winter depression is a profound phenomenon https://www.sciencedirect.com/science/article/abs/pii/0165032795000909.

2. Forecasting

2.1. SARIMA modeling

▶ Perform the Dickey-Fuller test. What is your conclusion?

```
adf.test(data$depression)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data$depression
## Dickey-Fuller = -3.2511, Lag order = 5, p-value = 0.08248
## alternative hypothesis: stationary
```

Since the p-value is bigger than 0.05 and if we consider the confidence interval of 95 percent, I believe that I cannot reject the null hypothesis. Thus, I cannot say that the data are stationary.

▶ Fit an (S)ARIMA model to the data; what is the order of the model that was selected?

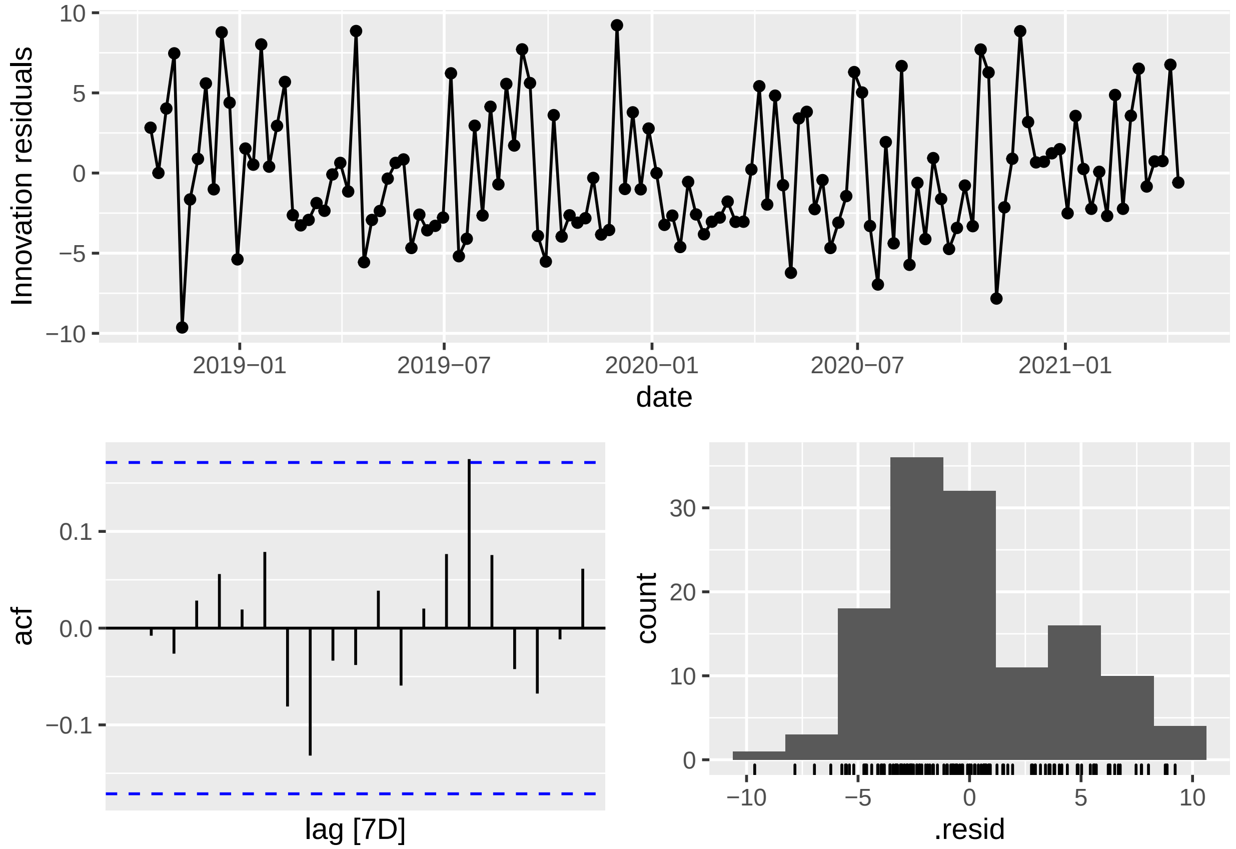
```
fit.d <- data %>% model(ARIMA(depression))
report(fit.d)
```

```
## Series: depression
## Model: ARIMA(2,0,0) w/ mean
##
## Coefficients:
##         ar1     ar2  constant
##       0.268  0.2578    2.7799
## s.e.  0.084  0.0850    0.3469
##
## sigma^2 estimated as 16.5:  log likelihood=-368.13
## AIC=744.27   AICc=744.59   BIC=755.77
```

After running the arima we can conclude that there was not needed to account for the seasonality since it already picked the best model. The degree of first differencing is zero together with the order of the moving average part. Since p = 2 the order of the autoregressive part is 2. This suggests that we have just an autoregression model. Since the constant c does not equal zero and d = 0, the long-term forecast will go to the mean of the data.

▶ Check the residuals of the model using the function gg_tsresiduals(). What is your conclusion?

```
gg_tsresiduals(fit.d)
```



From the sequence plot, I can see that the mean and variance do not change over time. Thus, it shows that the data are stationary. The autocorrelation is accounted for, as the correlations of lags are not significant and are around zero. This suggests that there is no structure in the residuals left. The histogram is skewed to the left side.

2.2. Dynamic regression

▶ Include the predictor in a dynamic regression model (i.e., allow for (S)ARIMA residuals); what is the effect of the predictor?

```
fit2 <- data %>% model(ARIMA(depression ~ mania))
report(fit2)
```

```
## Series: depression
## Model: LM w/ ARIMA(0,1,1) errors
##
## Coefficients:
##           ma1    mania
##       -0.8996  -0.9889
## s.e.   0.0436   0.0876
##
```

```
## sigma^2 estimated as 9.299:  log likelihood=-329.22
## AIC=664.45   AICc=664.64   BIC=673.05
```

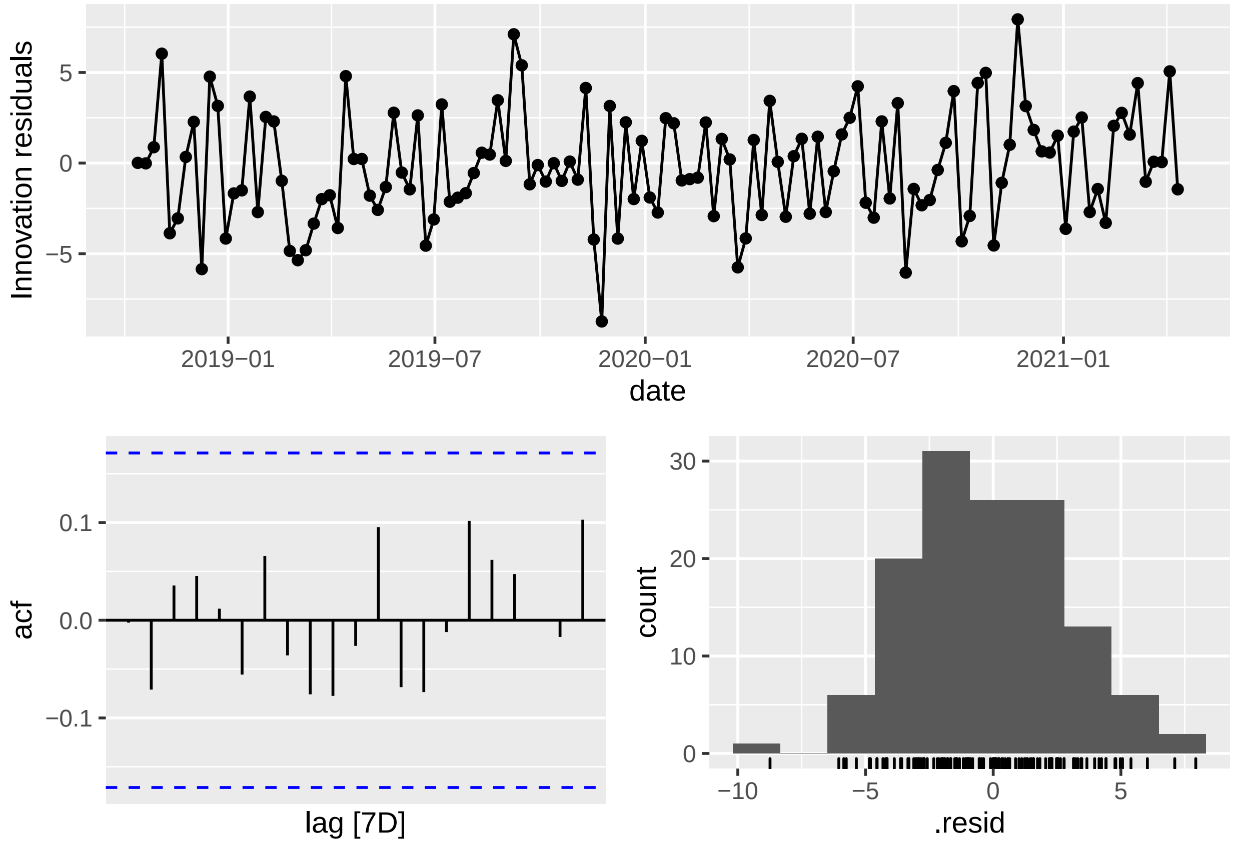
The effect of the predictor is negative. This means that the patient feels more depressed when he/she feels less in a mania state which makes sense since there is a fluctuation between mania, depression, and remission considering bipolar people. However, we should note that the two indicators come from one questionnaire but from different questions. This might bias the results. The effect of sleep median is also negative, but very small compared to the effect of mania.

▶ What order is the (S)ARIMA model for the residuals?

There is a need for differencing. Thus, d = 1. In this case, the p equals zero and q equals one which suggests a combination of moving average and random walk. Moreover, since the constant is zero and d = 1, the long-term forecasts will go to a non-zero constant.

▶ Check the residuals of the model using the function gg_tsresiduals(). What is your conclusion?

```
gg_tsresiduals(fit2)
```



The sequence plot looks to have a constant mean and constant variance. Thus, the data seem to be stationary. The histogram plot is skewed a little bit to the left and autocorrelations are close to zero and all seem to be rather insignificant. The estimated ARIMA errors are not significantly different from white noise.

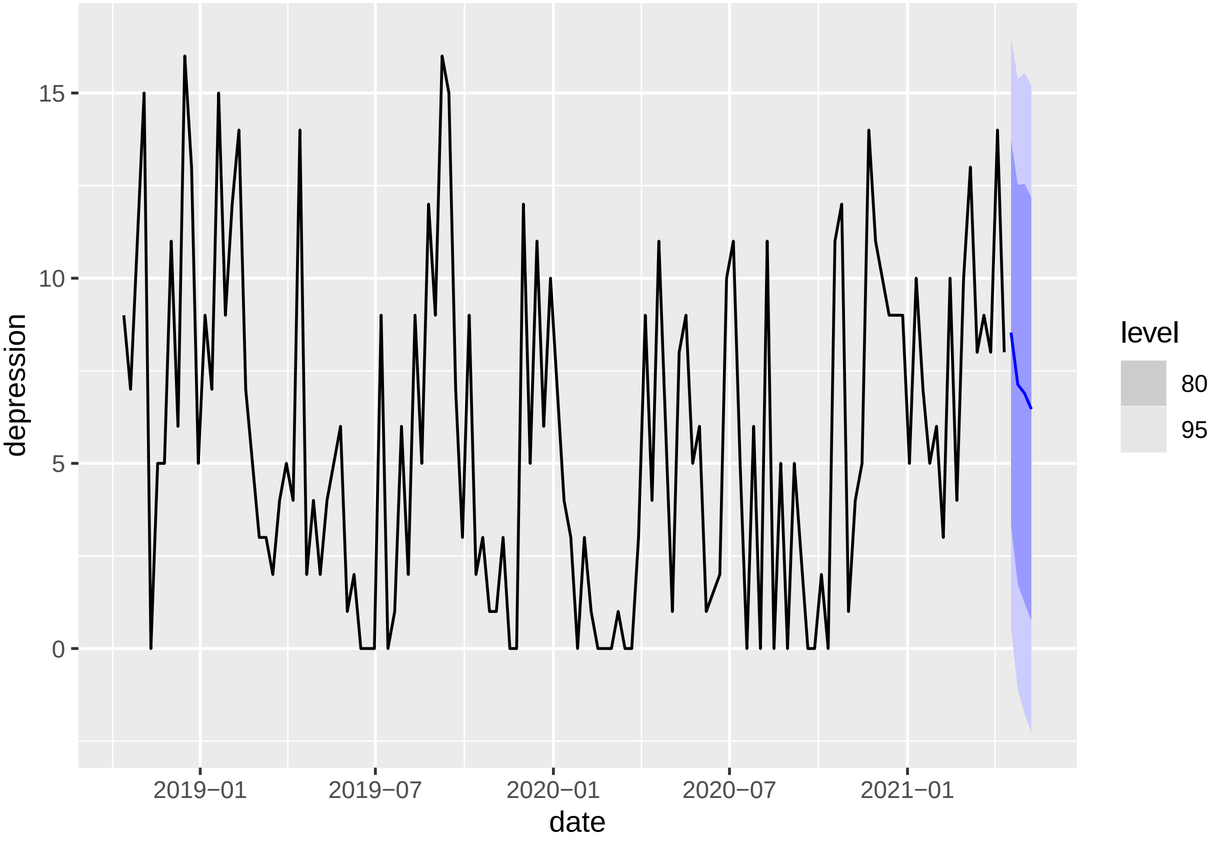
2.3. Forecasts

▶ Choose a forecasting horizon and indicate why this is a reasonable and interesting horizon to consider.

I believe that the interesting forecasting horizon is to look into the next four weeks. Longer horizons do not make sense since the data are very stationary. From the parameters and final equation, I think that I will always get lower and lower values until the mean. Thus, it is not feasible to consider longer horizons.

▶ Create forecasts based on the model without the predictor and plot these.

```
fit.d %>% forecast(h=4) %>% autoplot(data)
```



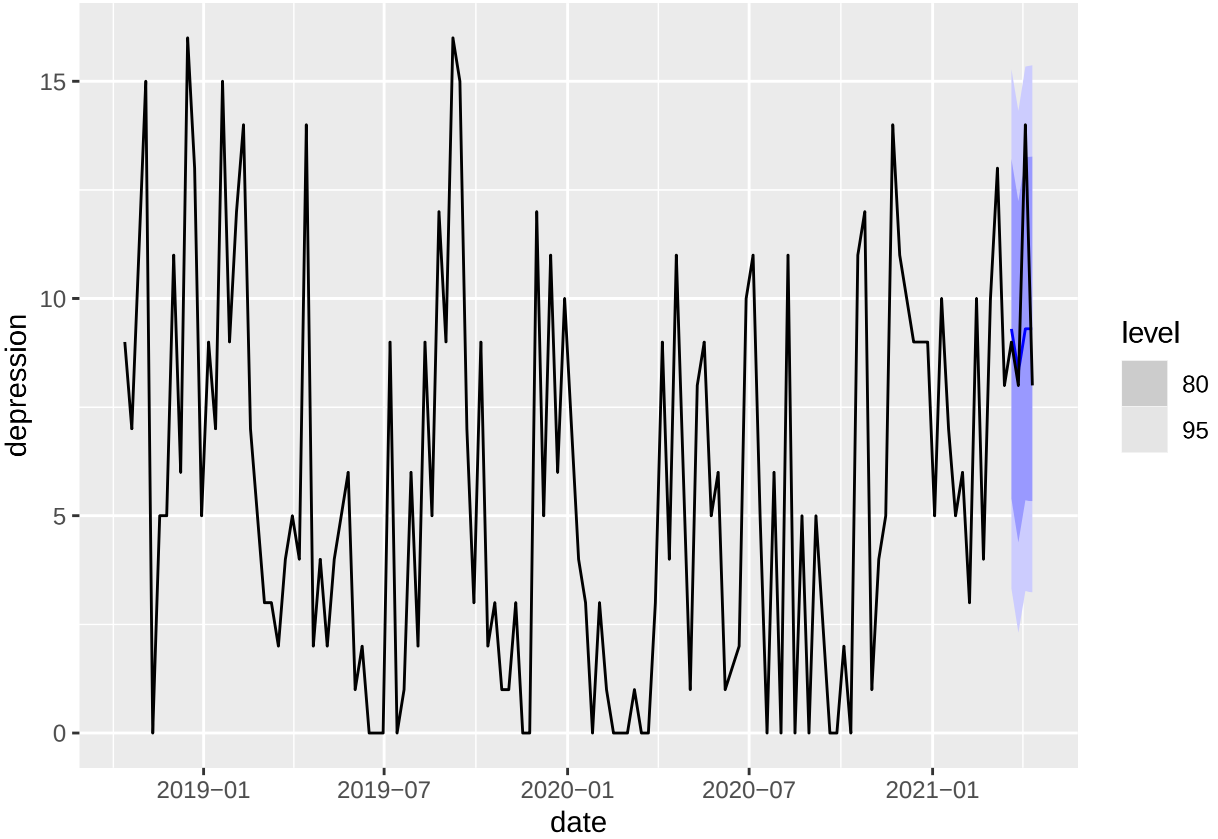
▶ Create forecasts based on the model with the predictor and plot these.

I will also include here the real data to see the comparison between predicted and real values!

```
temp_data <- tail(data, 4)
temp_data
```

```
## # A tsibble: 4 x 11 [7D]
##    date       mania depression actipoints_median actipoints_min actipoints_max
##    <date>     <dbl>      <dbl>             <dbl>          <dbl>          <dbl>
## 1 2021-03-21      0          9               990            895           1205
## 2 2021-03-28      1          8               970            785           1225
## 3 2021-04-04      0         14              1030            720           1110
## 4 2021-04-11      0          8               935            605           1035
## # ... with 5 more variables: actipoints_diff_mean <dbl>, sleep_median <dbl>,
## #   sleep_min <dbl>, sleep_max <dbl>, sleeps_diff_mean <dbl>
```

```
forecast(fit2, new_data = temp_data) %>% autoplot(data)
```



▶ Compare the plots of both forecasts (visually), and discuss how they are similar and/or different.

The confidence intervals are broader on the first plot without a mania predictor. Moreover, there seems to be a trend going down into the mean in the first plot. On the second plot, we can see some variation around the mean and more precise results.

3. Causal Modeling

▶ Formulate a causal research question(s) involving the time series variable(s) you have measured.

Considering the data it seems that this patient has very low periods of sleep compared to the general population. Sometimes bigger depression levels or depression episodes are associated with higher sleep. However, in this case, the mean of sleeping is 6,4 which is very small. Thus, the patient with a more regular regime and longer sleep could achieve better stability in terms of his/her bipolar disorder and depression levels.

So my question is: Does the lower sleep throughout the weeks causes bigger depression levels?

```
mean(data$sleep_median)
```

```
## [1] 6.441953
```

▶ Which method we learned about in class (Granger causal approaches, interrupted time series, synthetic controls) is most appropriate to answer your research question using the data you have available? Why?

The best and only method is the Granger causal approach because we do not have any intervention indicated in the data. Intervention in our data for example could be taking some medications such as lithium which is used commonly for treating bipolar people who are lithium respondents.

3.2 Analysis

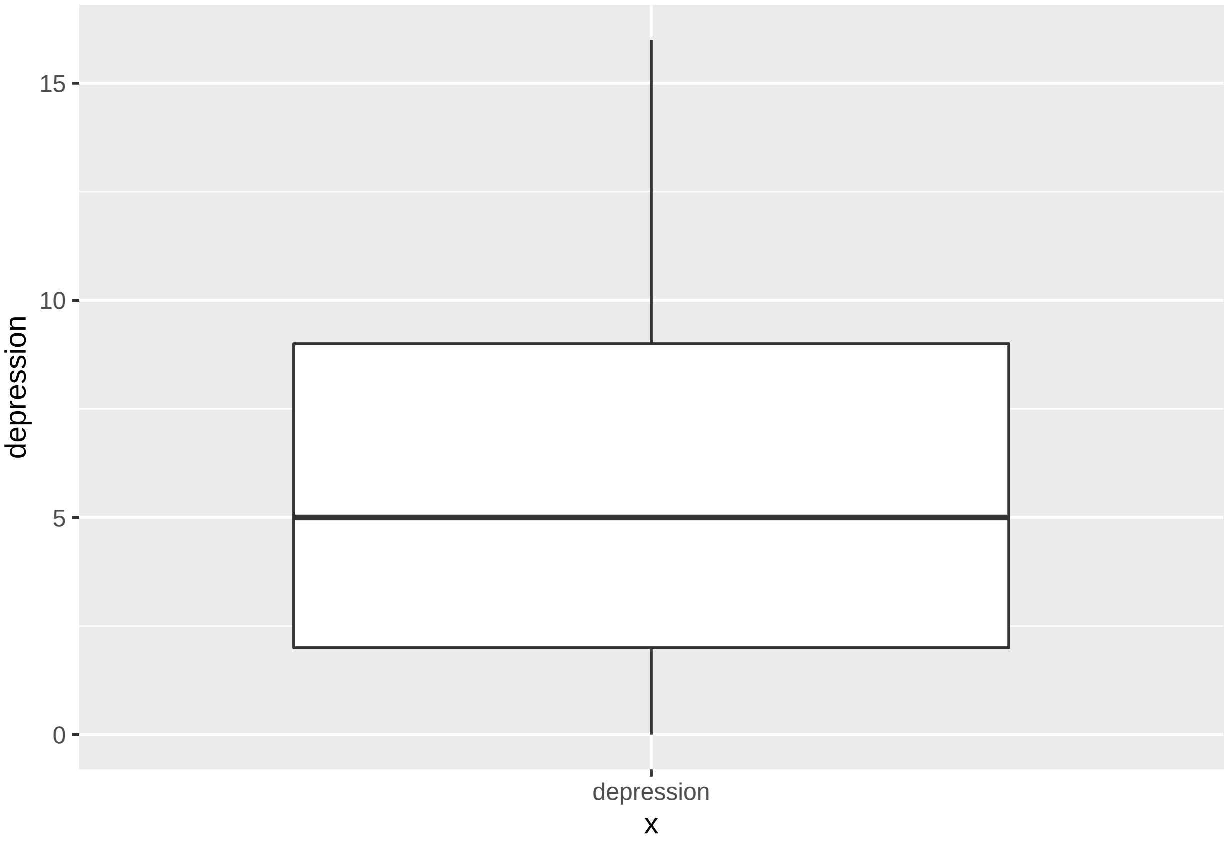
Depending on the choice you made above, follow the questions outlined in 3.2a, 3.2b, or 3.2c. If you chose a Granger causal analysis, it is sufficient to assess Granger causality in one direction only: you may evaluate a reciprocal causal relationship, but then answer each question below for both models.

I will answer the granger causal analysis just in one way due to the nature of the causal question.

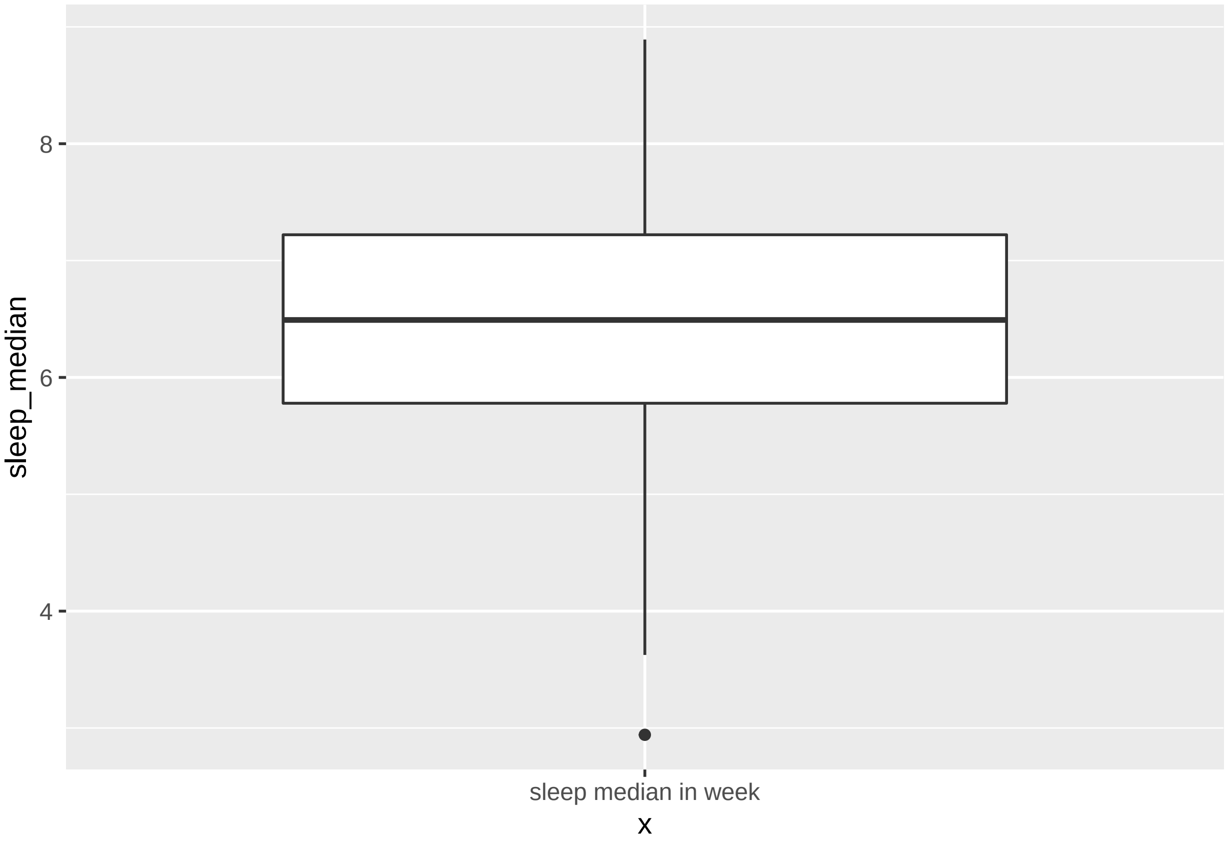
3.2a Granger Causal analysis

▶ Visualize your putative cause variable(s) X and outcome variables Y.

```
library(ggplot2)
ggplot(data, aes(x="depression", y=depression)) + geom_boxplot()
```

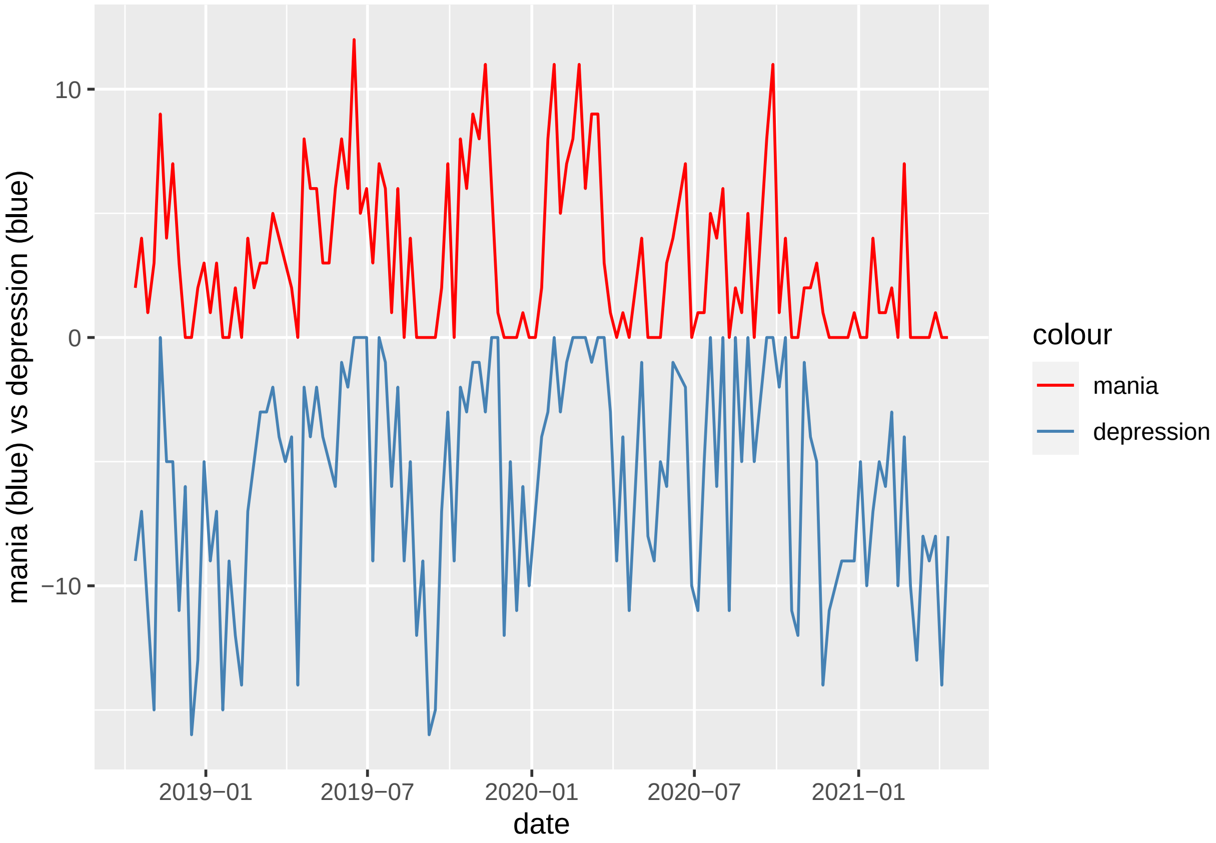


```
ggplot(data, aes(x="sleep median in week", y=sleep_median)) + geom_boxplot()
```



Here we can see that the patient sleeps less than 8 hours which could be considered as the population norm. Moreover the sleep is has a big variance from nearly 4 to 10 hours with one day - outlier when the patient was sleeping approximately two hours.

```
ggplot(data) + geom_line(  aes(x = date, y= - depression, color="steelblue")) + geom_lin
```

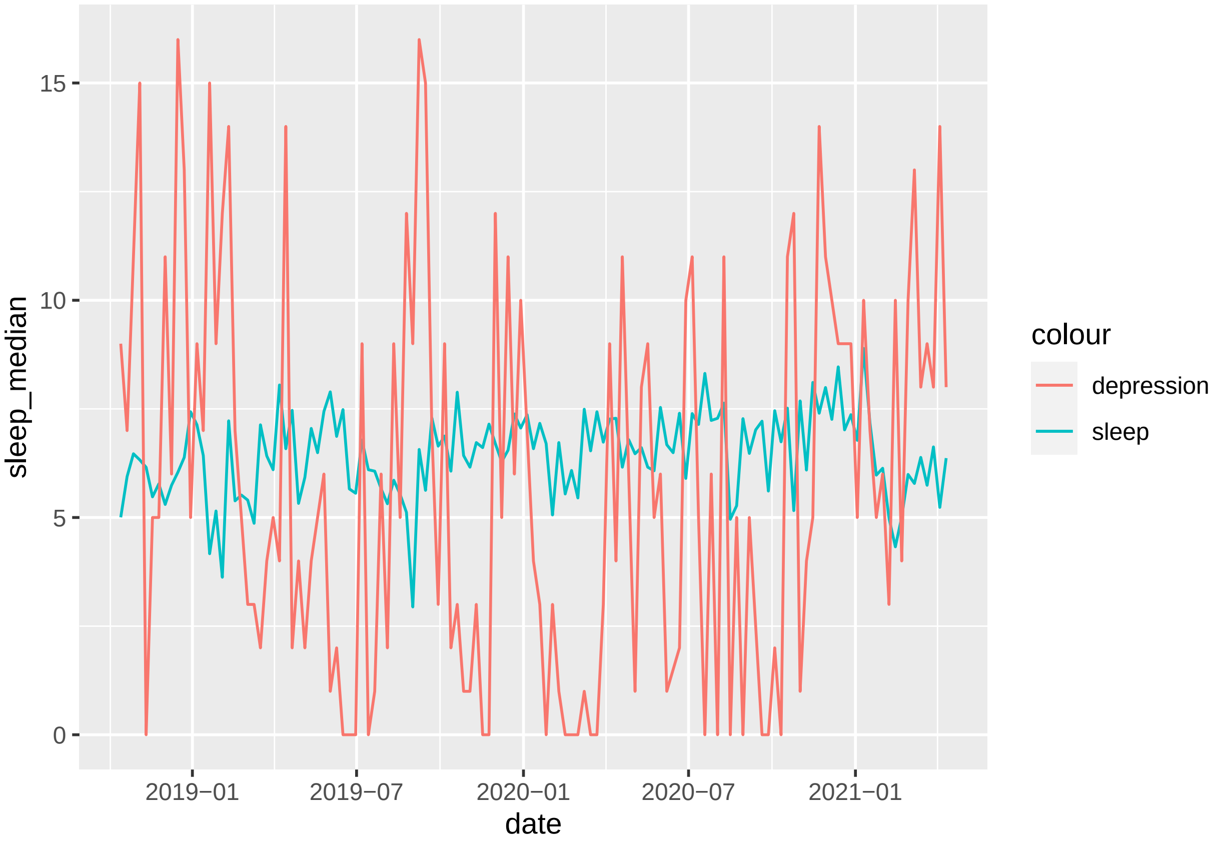


```
cor(data$depression, data$mania)
```

```
## [1] -0.7164223
```

Above we can see results of the mania questionnaire and depression below where mania is red and depression is blue.

```
ggplot(data) + geom_line(aes(x = date, y = sleep_median, color = "sleep")) + geom_line(a
```



Looking at the plot above, we can cannot see any specific trend between sleep and depression.

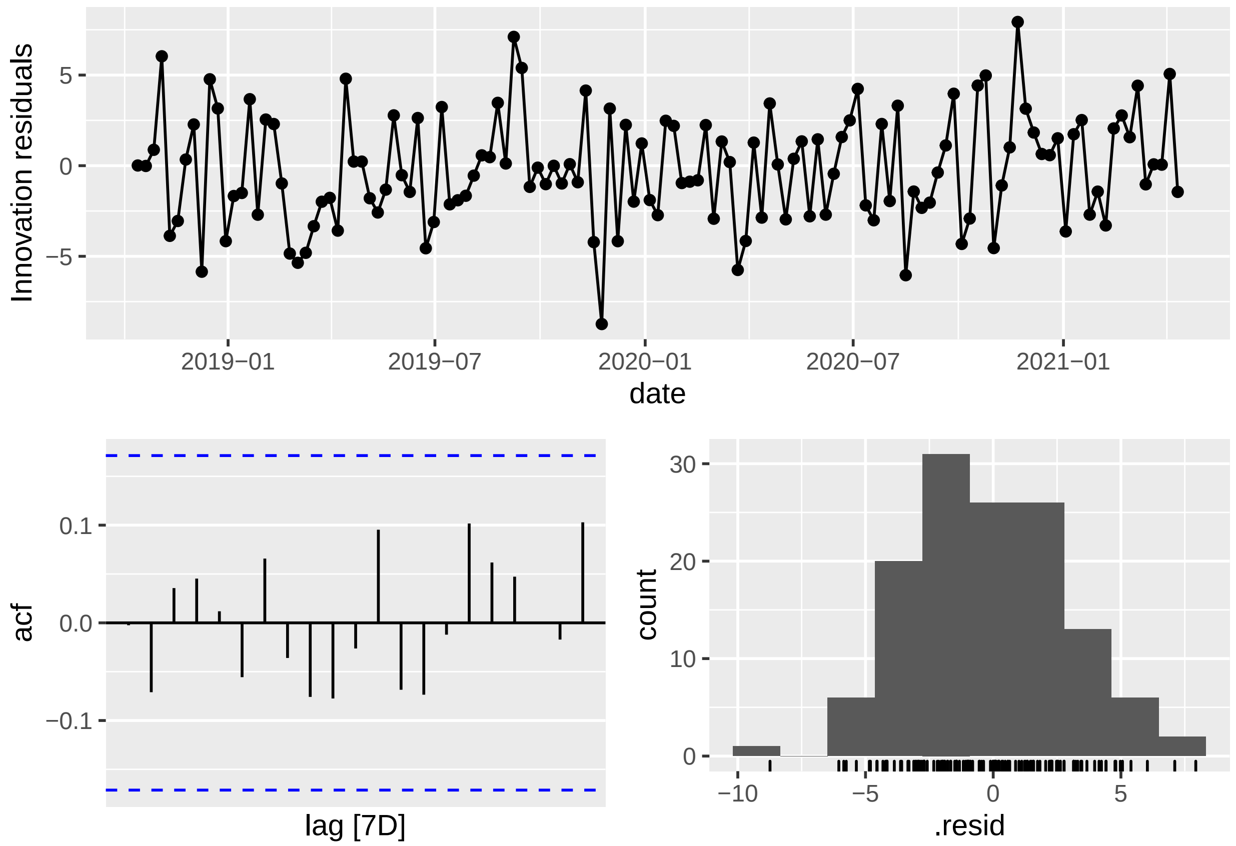
▶ Train an appropriate ARIMA model on your outcome variable(s) Y, ignoring the putative cause variable(s) (X) but including, if appropriate, any additional covariates. If using the same model as fit in part 2, briefly describe that model again here.

```
fit2 <- data %>% model(ARIMA(depression ~ mania))
report(fit2)
```

```
## Series: depression
## Model: LM w/ ARIMA(0,1,1) errors
##
## Coefficients:
##           ma1    mania
##       -0.8996  -0.9889
## s.e.   0.0436   0.0876
##
## sigma^2 estimated as 9.299:  log likelihood=-329.22
## AIC=664.45   AICc=664.64   BIC=673.05
```

This model takes into account mania levels when determining depression. There is a need for differencing. Thus, d = 1. In this case, the p equals zero and q equals one which suggests a combination of moving average and random walk. Moreover, since the constant is zero and d = 1, the long-term forecasts will go to a non-zero constant.

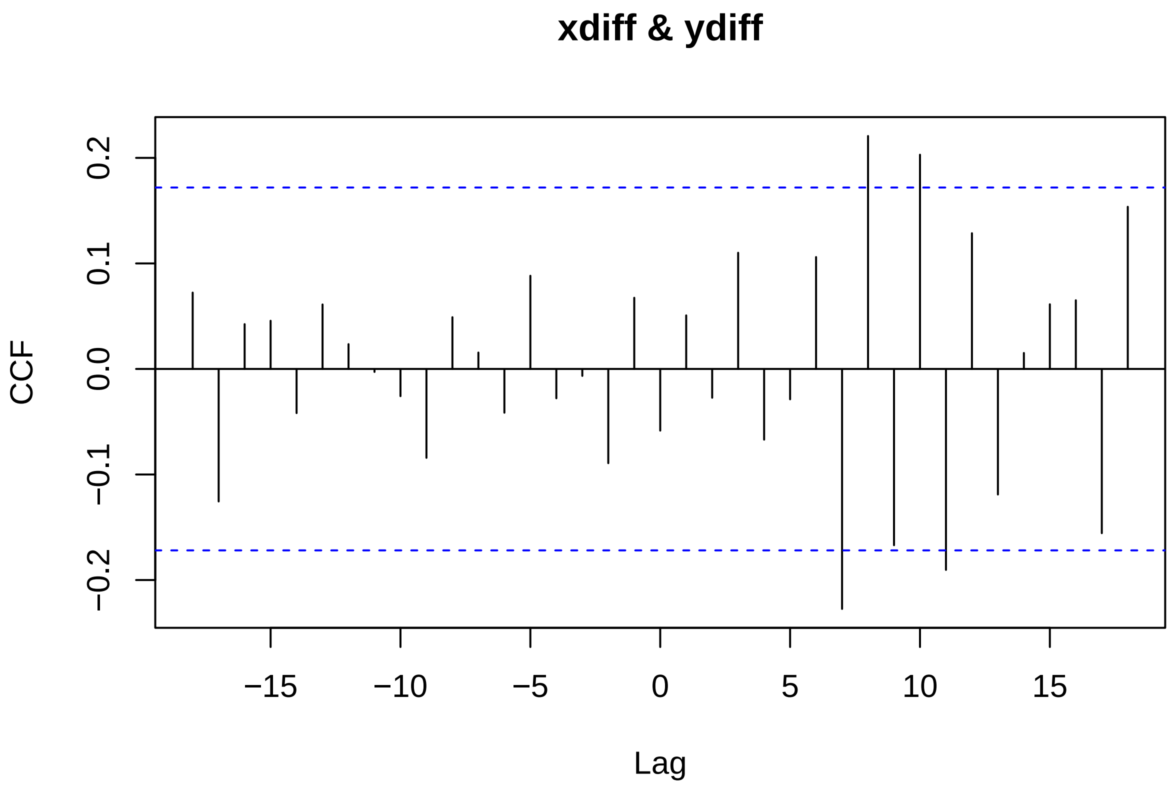
```
gg_tsresiduals(fit2)
```



When looking at the residuals we can see that the differencing worked and that the sequence plots have the same mean and variance. Thus, the data are stationary. The histogram is a little bit skewed.

▶ Justify what range of lags to consider for the lagged predictor(s). Use the CCF, but you may also justify this based on domain knowledge or substantive theory.

```
xdiff <- diff(data$sleep_median, lag=1)
ydiff <- diff(data$depression, lag = 1)
ccf(xdiff, ydiff, ylab = "CCF")
```



According to the plot, the meaningful (even though very small) cross-correlation is at lag 7 to 11 - instead of number 9. Thus, I will run multiple models where I will use consecutive correlation until lag of 11.

▶ Investigate whether adding your lagged "cause'' variables (X) improve the prediction of your effect variable(s) Y. Use model selection based on information criteria. Describe your final chosen model

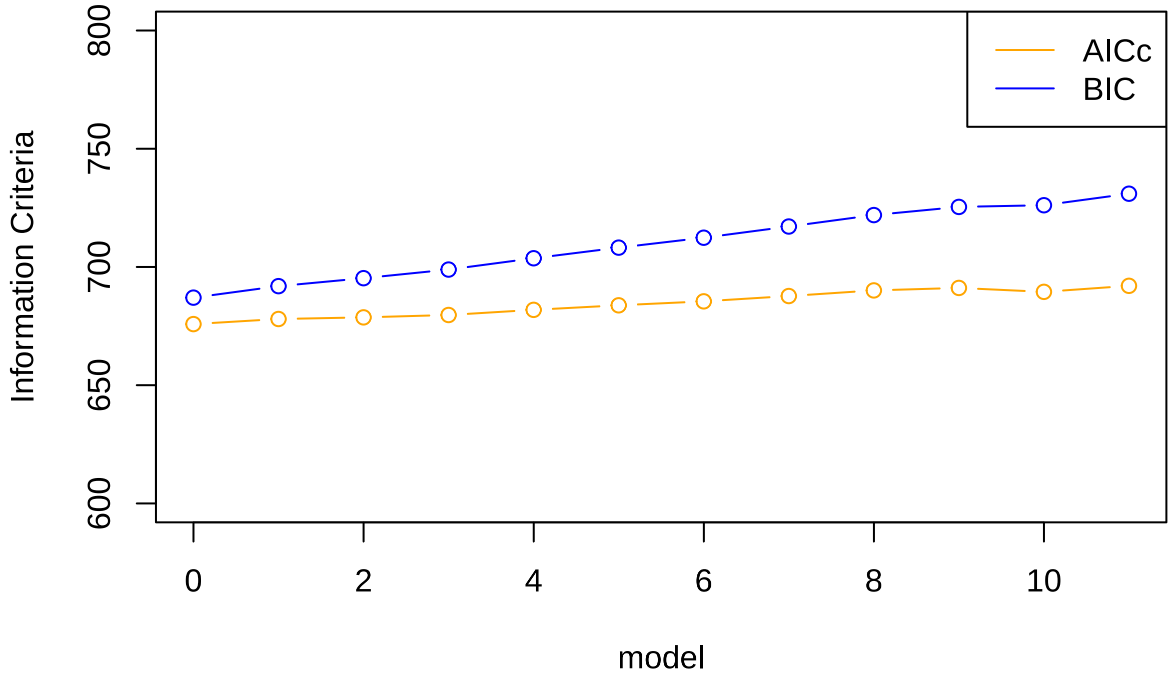
```
data$x = data$sleep_median
data$y = data$depression
 fit <- data %>%
  # Restrict data so models use same fitting period
  mutate(y = c(NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, y[12:131])) %>%
  # Estimate models
  model(
    indep = ARIMA(y),
    lag1 = ARIMA(y~ lag(x)),
    lag2 = ARIMA(y~ lag(x) + lag(x,2)),
    lag3 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3)),
    lag4 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4)),
    lag5 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5)),
    lag6 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6)),
    lag7 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6) +  lag
    lag8 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6) + lag(
    lag9 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6) + lag(
    lag10 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6) + lag
    lag11 = ARIMA(y~ lag(x) + lag(x,2) + lag(x,3) + lag(x,4) + lag(x,5) + lag(x,6) + lag

  )

glance(fit)
```

```
## # A tibble: 12 x 8
##    .model sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##    <chr>   <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
##  1 indep    14.3   -334.  676.  676.  687. <cpl [0]> <cpl [2]>
##  2 lag1     14.4   -334.  678.  678.  692. <cpl [0]> <cpl [2]>
##  3 lag2     14.3   -333.  678.  679.  695. <cpl [0]> <cpl [2]>
##  4 lag3     14.3   -332.  679.  680.  699. <cpl [0]> <cpl [2]>
##  5 lag4     14.4   -332.  681.  682.  704. <cpl [0]> <cpl [2]>
##  6 lag5     14.5   -332.  682.  684.  708. <cpl [0]> <cpl [2]>
##  7 lag6     14.5   -332.  684.  685.  712. <cpl [0]> <cpl [2]>
##  8 lag7     14.6   -332.  685.  688.  717. <cpl [0]> <cpl [2]>
##  9 lag8     14.7   -332.  687.  690.  722. <cpl [0]> <cpl [2]>
## 10 lag9     14.7   -331.  688.  691.  725. <cpl [0]> <cpl [2]>
## 11 lag10    14.3   -329.  686.  689.  726. <cpl [2]> <cpl [0]>
## 12 lag11    14.4   -329.  688.  692.  731. <cpl [2]> <cpl [0]>
```

```
plot(seq(0,11), glance(fit)$AICc,
     col = "orange", type = "b",
     ylab = "Information Criteria", xlab = "model",
     ylim = c(600,800))
lines(seq(0,11), glance(fit)$BIC, col = "blue", type = "b")
legend("topright", c("AICc","BIC"), col = c("orange","blue"), lty = 1)
```



The final chosen model is the baseline model which does not take into account the sleep variable since it has the lowest AIC and simultaneously the BIC value. I already described this model but to remind myself: the degree of first differencing is zero together with the order of the moving average part. Since p = 2 the order of the autoregressive part is 2. This suggests that we have just an autoregression model. Since the constant c does not equal zero and d = 0, the long-term forecast will go to the mean of the data.

```
fit_best_aic <- data %>% model(ARIMA(y))
report(fit_best_aic)
```

```
## Series: y
## Model: ARIMA(2,0,0) w/ mean
##
## Coefficients:
##          ar1     ar2  constant
##        0.268  0.2578    2.7799
## s.e.   0.084  0.0850    0.3469
##
## sigma^2 estimated as 16.5:  log likelihood=-368.13
## AIC=744.27   AICc=744.59   BIC=755.77
```

3.3 Conclusion and critical reflection

▶ Based on the result of your analysis, how would you answer your causal research question?

The sleep_median per week would be a granger cause of depression if the AIC or BIC would be lower than the base model. However, this is not the case as all the models with a lagged version of sleep_median have greater AIC and BIC than the baseline model without the sleep_median information. Thus, we can conclude that sleep_median per week is not a granger-cause of depression in this case.

▶ Making causal conclusions on the basis of your analysis is reliant on a number of assumptions. Pick a single assumption that is necessary in the approach you chose. Discuss the plausability and possible threats to the validity of this assumption in your specific setting (< 75 words)

As my conclusion was that sleep is not a granger cause of depression, we would have to assume that there are no unobserved confounding. However, the person's sleep and simultaneously the depression can be dependent on multiple other things in life. Thus, this is a very possible threat to my analysis.
