# Met4 Home Exam Spring 2020 Sensor guide

## Introduction

This document is a guide to how the home exam in Met4 can be solved. It is not, however, a solution manual as there are many ways the case can be solved.

We start by loading in the data and required libraries:

```
rm(list=ls())

# Load necessary libraries:
library(tidyverse)
library(patchwork)
library(magrittr)
library(rlang)

# Load the avinor-data.
load("avinor_cleaned.Rdata")

df_total %<>% filter(dep_date < Sys.Date())
df_airline %<>% filter(dep_date < Sys.Date())
```

## Q1

We are asked to provide some summaries of the data. Here, we supply some simple descriptives of all the variables in the df_total dataset. The most interesting variables in this data set are flights - i.e. the number of flights per day, and cumulative daily flights. These variables are provided as a total, as well as split by international and domestic flights. In addition to the flights data we have a date, the day of the week, and a trend variable that simply increases by 1 by each date.

```
df_total %>%
  summary() %>%
  knitr::kable()
```

| dep_date | flights | flights_int | flights_dom | cumflights | cumflights_int | cumflights_dom | day_of_week | trend |
|---|---|---|---|---|---|---|---|---|
| Min. :2020-01-20 | Min. : 63.0 | Min. : 9.0 | Min. : 52.0 | Min. : 914 | Min. : 225 | Min. : 689 | 1:11 | Min. : 1 |
| 1st Qu.:2020-02-08 | 1st Qu.:372.0 | 1st Qu.: 98.0 | 1st Qu.:276.0 | 1st Qu.:16181 | 1st Qu.: 4275 | 1st Qu.:11906 | 2:11 | 1st Qu.:20 |
| Median :2020-02-27 | Median :835.0 | Median :216.0 | Median :619.0 | Median :31860 | Median : 8600 | Median :23260 | 3:11 | Median :39 |
| Mean :2020-02-27 | Mean :672.6 | Mean :172.1 | Mean :500.5 | Mean :30382 | Mean : 8105 | Mean :22277 | 4:11 | Mean :39 |
| 3rd Qu.:2020-03-17 | 3rd Qu.:924.0 | 3rd Qu.:234.0 | 3rd Qu.:691.0 | 3rd Qu.:46724 | 3rd Qu.:12673 | 3rd Qu.:34051 | 5:11 | 3rd Qu.:58 |
| Max. :2020-04-05 | Max. :999.0 | Max. :256.0 | Max. :754.0 | Max. :51791 | Max. :13249 | Max. :38542 | 6:11 | Max. :77 |

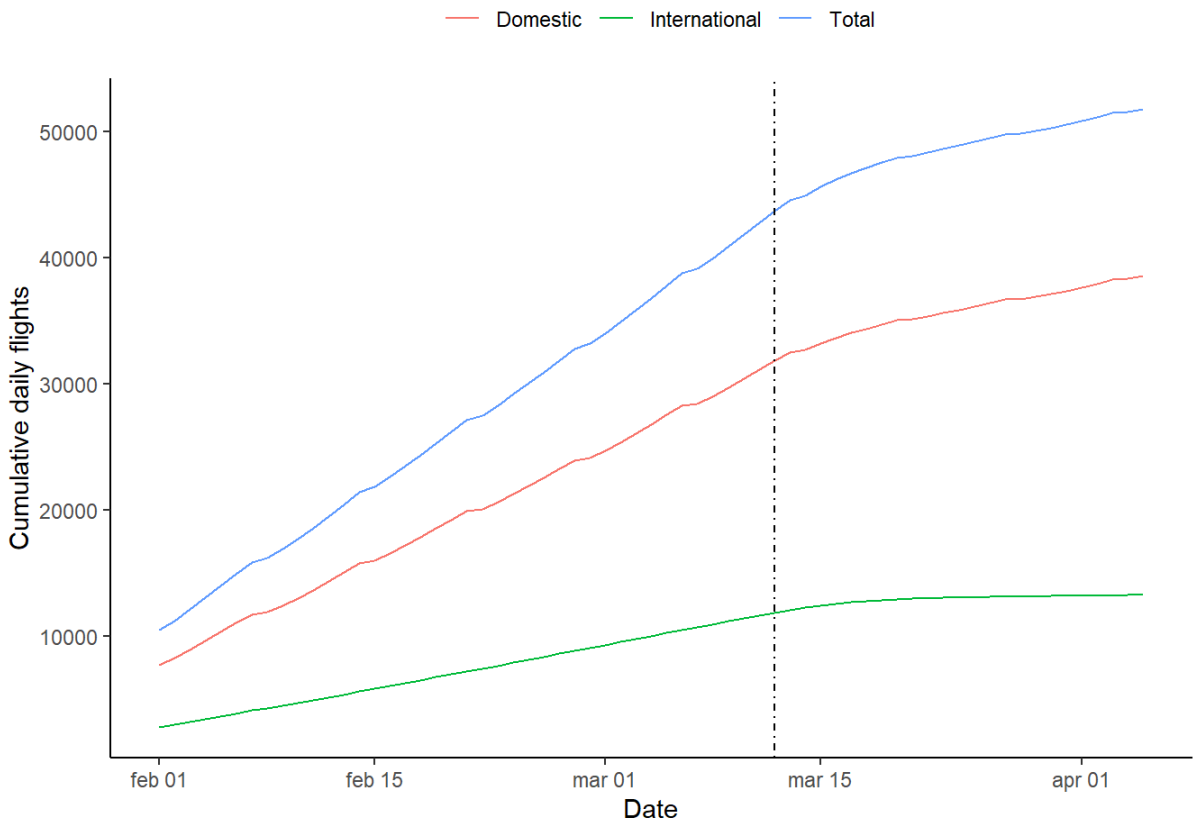
| dep_date | flights | flights_int | flights_dom | cumflights | cumflights_int | cumflights_dom | day_of_week | trend |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | 7:11 | |

This is only a summary of the summarised data set. We are also given a raw data set where each observation is an individual flight. This raw data set contains significantly more information (airfield, gate number, whether the flight was delayed etc). The question is not clear exactly which data set to provide summaries for. The most important part is to provide summaries that are relevant for the remainder of the assignment, so which variables should be included in this assignment depends on how question 4 is solved.

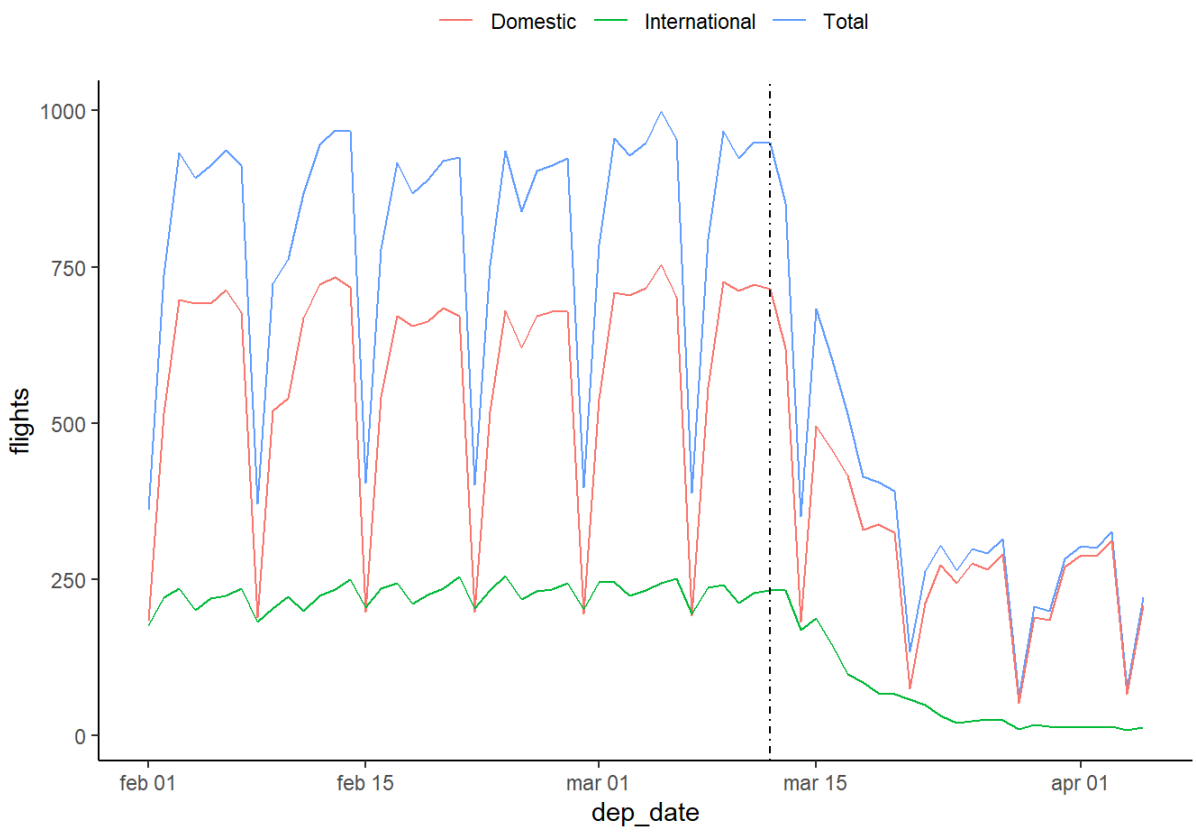
Further, we are asked to decide on which date the corona crisis became visible in departures. To answer this we need to inspect the data. This can best be achieved with plots. Below, we plot cumulative flights and daily flights after February, and the idea is simply to inspect the figures and “eyeball” where we see a drop (of course, more advanced methods exist for detecting a cutoff date).

From visual inspection, it appears that the daily flights started dropping significantly after March 13. We can also not that there clearly are seasonalities in the data, where there number of flights vary through the week. It could also be some seasonalities with lower frequency, but these are hard to detect when we only have a few weeks of data.

```
min_plot_date <- as.Date('2020-02-01')
cutoff_date <- as.Date('2020-03-12')

df_total %>%
  filter(dep_date >= min_plot_date) %>%
  ggplot(aes(x = dep_date)) +
  geom_line(aes(y = cumflights, col = "Total")) +
  geom_line(aes(y = cumflights_dom, col = "Domestic")) +
  geom_line(aes(y = cumflights_int, col = "International")) +
  geom_vline(xintercept = as.numeric(cutoff_date), linetype = 4) +
  ylab("Cumulative daily flights") +
  xlab("Date") +
  labs(col = "") +
  theme_classic() +
  theme(legend.position = "top")
```



```
df_total %>%
  filter(dep_date >= min_plot_date) %>%
  filter(dep_date < Sys.Date()) %>%
  ggplot(aes(x = dep_date)) +
  geom_line(aes(y = flights, col = "Total")) +
  geom_line(aes(y = flights_dom, col = "Domestic")) +
  geom_line(aes(y = flights_int, col = "International")) +
  geom_vline(xintercept = as.numeric(cutoff_date), linetype = 4) +
  labs(col = "") +
  theme_classic() +
  theme(legend.position = "top")
```




# Q2

We next run regression on all the flight variables, using a trend and indicators for weekdays as covariates, and only data up until the cutoff date.

The $R^2$ is very high in all regressions. This is natural when there are strong seasonalities and trends in the data, as most of the variation in the dependent variable will be picked up by these covariates. The weekday variables are generally significant in all the regressions, with the exception of international flights. It could maybe be the case that vacation flights follow a differents pattern, and that winter break occured in the observed time period. The first day of the week is Sunday in this model, so we see a general pattern that there are fewer flights in the weekend.

The trend is significant in all but one regression, and particularly so in the cumulative flights. The interpretation of e.g. the trend coefficient in the cumulative total flights regression is that there are, on average, 815 flights every day at Avinor airports.

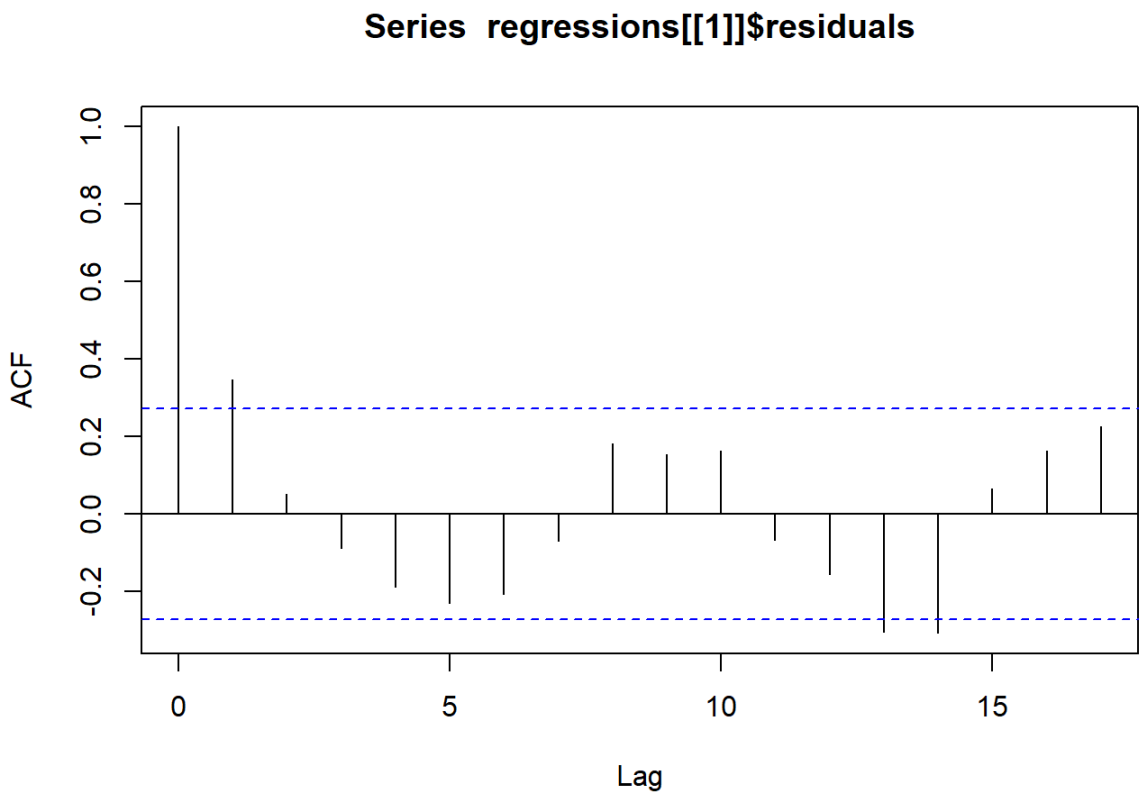
```
dep.vars <- colnames(df_total)[grep("flight",colnames(df_total))]
regressions <- vector("list", length(dep.vars))
predictions <- vector("list", length(dep.vars))

for(i in 1:length(dep.vars)){

  form <-
    formula(
      paste0(
        dep.vars[i],
        "~trend + day_of_week"))

  regressions[[i]] <-
    lm(form,
       data = df_total,
       subset = dep_date < cutoff_date)

  predictions[[i]] <-
    predict(regressions[[i]],
            newdata = df_total,
            interval = "prediction") %>%
    as.data.frame()
}

stargazer::stargazer(regressions,
                     type = "html")
```

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | flights | flights_int | flights_dom | cumflights | cumflights_int | cumflights_dom |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| trend | 1.072*** | 0.540*** | 0.532* | 814.955*** | 224.298*** | 590.656*** |
| | (0.301) | (0.066) | (0.279) | (1.124) | (0.666) | (0.698) |
| day_of_week2 | 158.019*** | 12.028*** | 145.990*** | 122.208* | 33.782 | 88.427** |
| | (16.818) | (3.681) | (15.616) | (62.905) | (37.261) | (39.043) |
| day_of_week3 | 124.197*** | -19.637*** | 143.833*** | 189.128*** | 16.608 | 172.520*** |
| | (16.807) | (3.678) | (15.606) | (62.865) | (37.237) | (39.018) |
| day_of_week4 | 156.125*** | -3.551 | 159.677*** | 289.049*** | 16.060 | 272.989*** |
| | (16.802) | (3.677) | (15.601) | (62.845) | (37.226) | (39.005) |
| day_of_week5 | 186.929*** | 5.048 | 181.881*** | 375.864*** | 10.466 | 365.398*** |
| | (17.376) | (3.803) | (16.133) | (64.991) | (38.497) | (40.337) |
| day_of_week6 | 170.429*** | 17.366*** | 153.064*** | 488.481*** | 30.025 | 458.456*** |
| | (17.363) | (3.800) | (16.121) | (64.942) | (38.468) | (40.307) |
| day_of_week7 | -376.071*** | -36.032*** | -340.040*** | 55.669 | -3.273 | 58.942 |
| | (17.355) | (3.798) | (16.114) | (64.913) | (38.451) | (40.289) |
| Constant | 729.277*** | 212.454*** | 516.824*** | -180.303*** | -158.639*** | -21.664 |
| | (14.880) | (3.256) | (13.816) | (55.654) | (32.966) | (34.542) |
| Observations | 52 | 52 | 52 | 52 | 52 | 52 |
| $R^2$ | 0.974 | 0.891 | 0.974 | 1.000 | 1.000 | 1.000 |
| Adjusted $R^2$ | 0.970 | 0.873 | 0.970 | 1.000 | 1.000 | 1.000 |
| Residual Std. Error (df = 44) | 32.463 | 7.105 | 30.142 | 121.422 | 71.924 | 75.363 |
| F Statistic (df = 7; 44) | 236.142*** | 51.188*** | 235.972*** | 75,324.730*** | 16,269.060*** | 102,704.000*** |
| *Note:* | | | | | | *p<0.1;* ***p<0.05;*** p<0.01 |

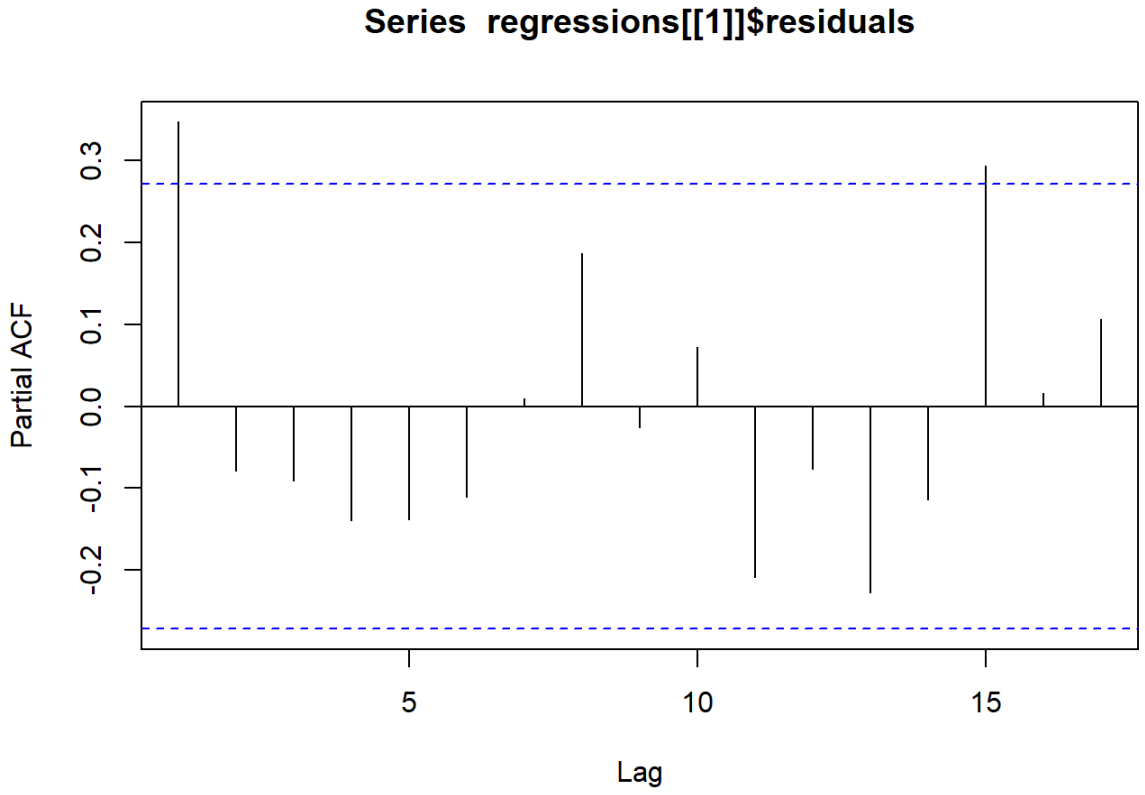
We can also plot some diagnostics. In general, the most worrying of these is the autocorrelation in the residuals of the cumulative flights regression. Strictly speaking, this dependence of the residuals makes the inference invalid. It is not surprising that this dependence is present, as e.g. an abnormally high flight count in a given day will be present also in future observations.

We could account for this dependence in a time series model. However, ensuring we have the correct significance level of the coefficients is not central to this assignment, so we'll leave it at that - although some students might choose to investigate this further in assignment 4.

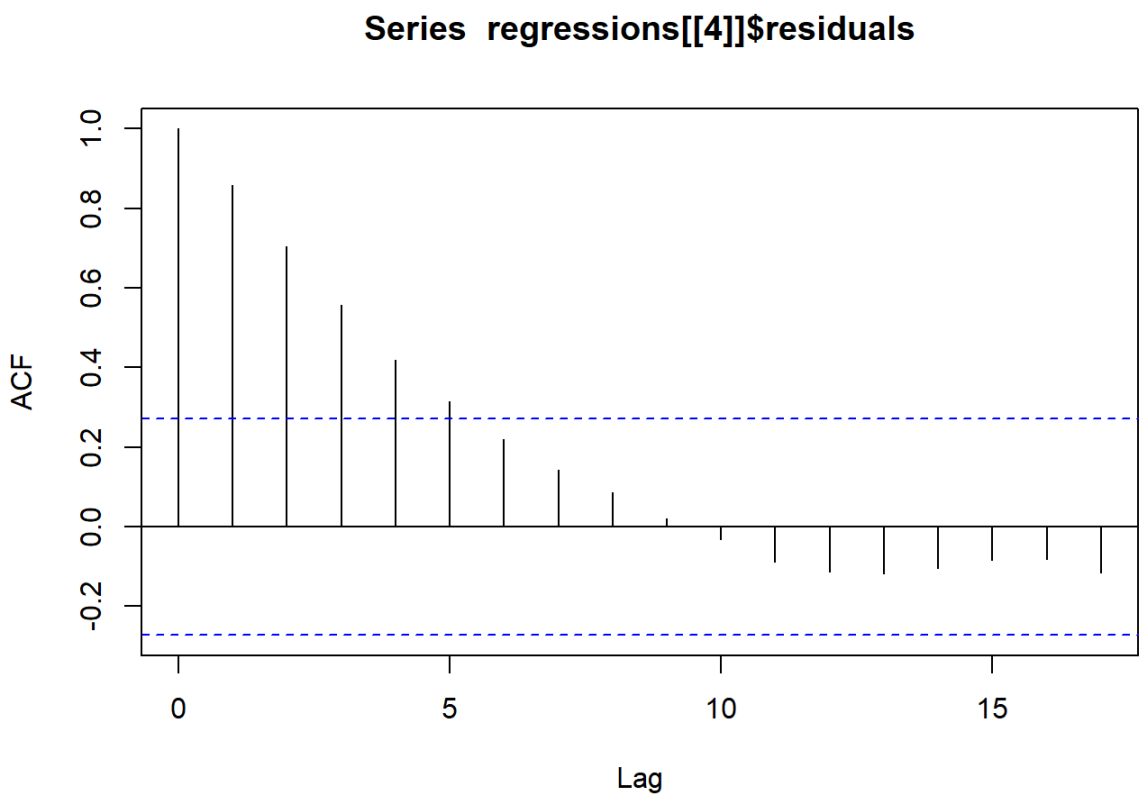
```
#plot(regressions[[1]])
acf(regressions[[1]]$residuals)
```



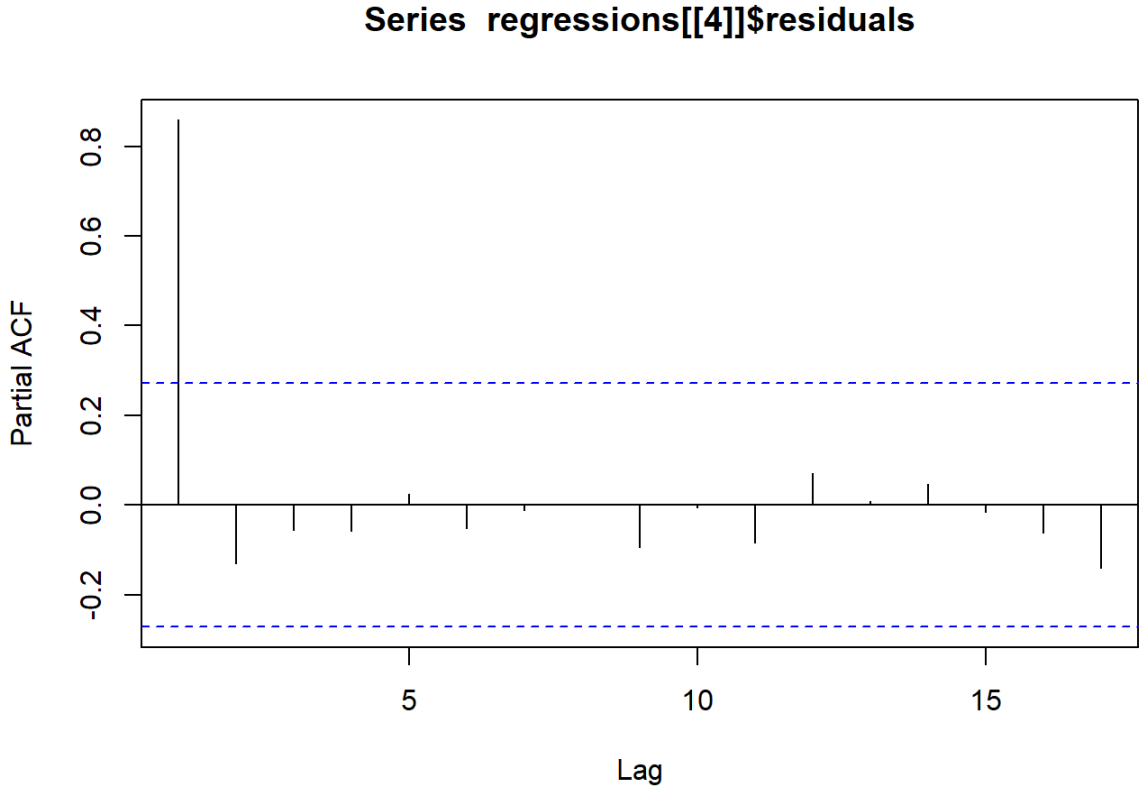

```
pacf(regressions[[1]]$residuals)
```




```
#plot(regressions[[4]])
acf(regressions[[4]]$residuals)
```



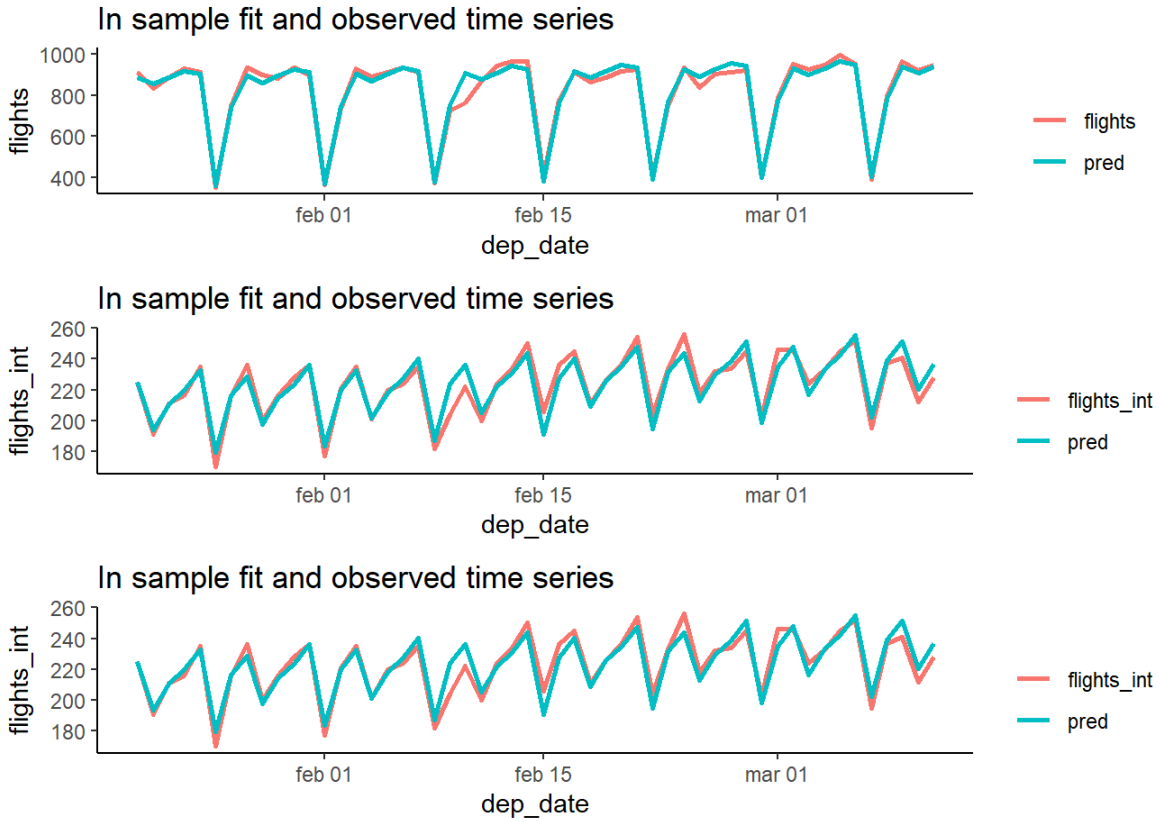

```
pacf(regressions[[4]]$residuals)
```



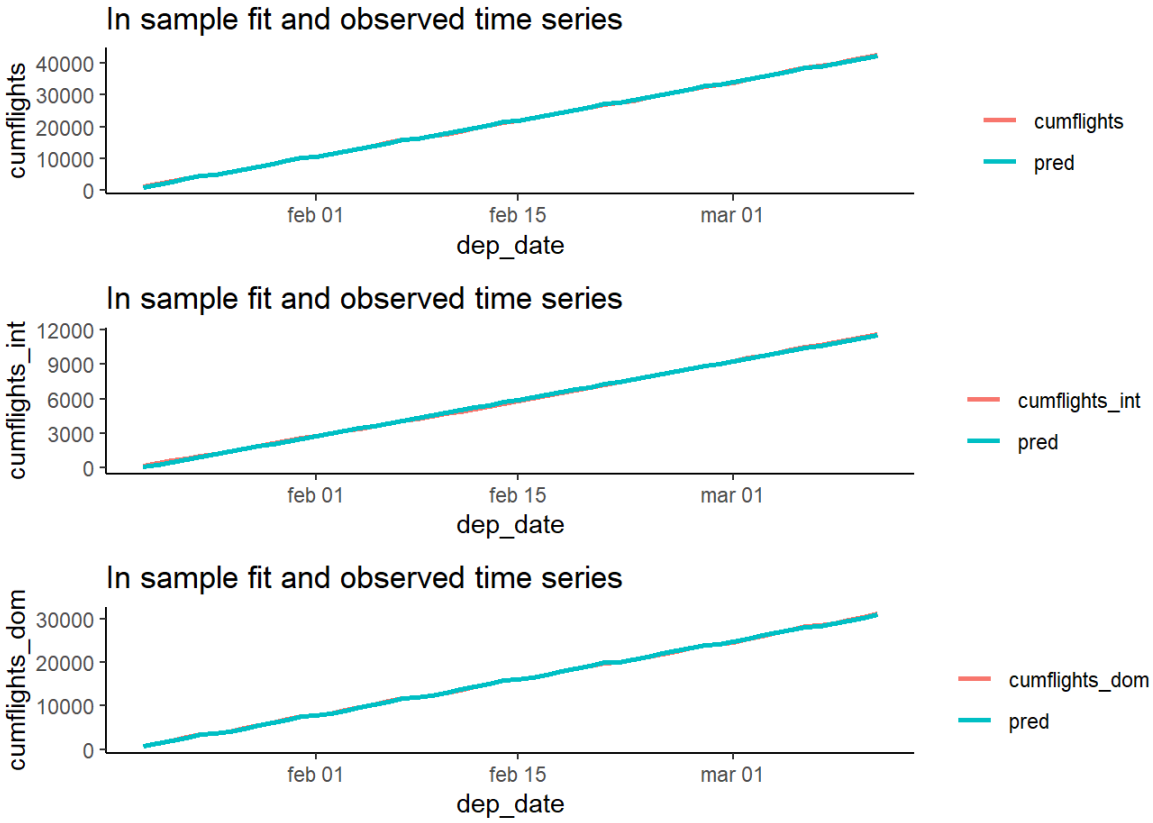

As a final check, we can plot the time series agains their in sample predictions. For both daily flights and cumulative flights we capture the time series very well with these simple models (which should be no surprise, given the high $R^2$). There is some noise around the daily flights vs. the in sample predictions, but for the most part the in sample prediction tracks the time series well. For the cumulative flights it is hard to even observe differences between the data and the model. Good submissions to the exam might want to try more serious methods for evaluating the fit of the model, using e.g. out of sample forecasts.

In all, the models seem to fit the data well, and can be used to assess the impact of the Corona crisis on flight departures. An important caveat is though that we are estimating the model on a short time period. If there are special occasions coming up (e.g. Easter) where departures normally are different from the patterns we observe from January to March, then this will not be accounted for by our model. Hence, using these models we are really only assessing whether traffic patterns in the future is different from what we observe in Jan-Mar.

```
plot.reg.in.sample <-
  function(i) {
    df_total %>%
      bind_cols(predictions[i]) %>%
      filter(dep_date < cutoff_date) %>%
      ggplot(aes(x = dep_date)) +
      geom_line(aes(y = get(dep.vars[i]), col = dep.vars[i]), lwd = 1) +
      geom_line(aes(y = fit, col = "pred"), lwd = 1) +
      ylab(dep.vars[i]) +
      theme_classic() +
      labs(col = "") +
      ggtitle("In sample fit and observed time series")
  }

(
  plot.reg.in.sample(1)
  / plot.reg.in.sample(2)
  / plot.reg.in.sample(2)
)
```



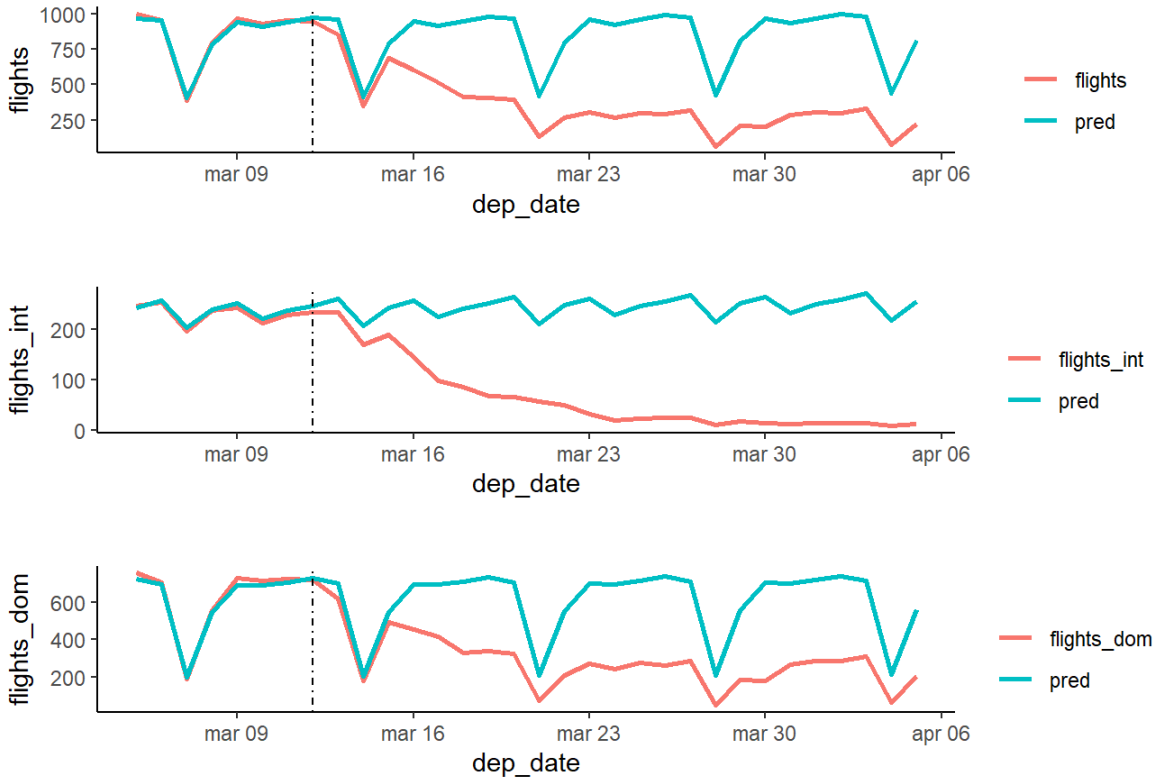

```
(
  plot.reg.in.sample(4)
  / plot.reg.in.sample(5)
  / plot.reg.in.sample(6)
)
```




# Q3

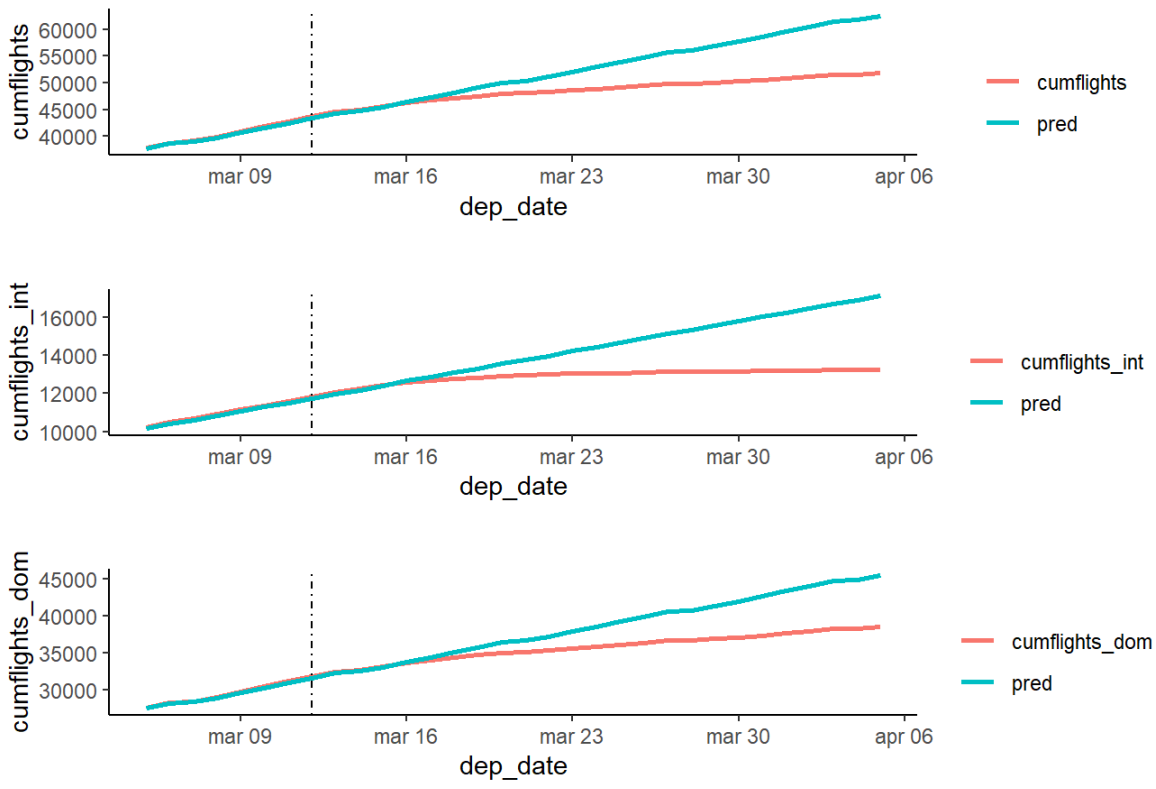
With the models estimated, all we have to do to quantify the effect of the corona crisis on flight departures is to compare the models' predictions against the observed time series *after* the cutoff date. First we can plot the time series themselves against the predictions. For the daily flights data, we can observe a sharp drop in departures versus the expectation immediately after the cutoff date. There is perhaps some tendency towards that international flights were starting to decrease slightly already before the cutoff date. Note that this document was written well before the exam - so there might happen more interesting things in the time series between the time this document is written and the exam.

```
plot.reg <-
  function(i){
    df_total %>%
      bind_cols(predictions[i]) %>%
      filter(dep_date >= cutoff_date - 7) %>%
      ggplot(aes(x=dep_date))+
      geom_line(aes(y=get(dep.vars[i]), col=dep.vars[i]), lwd=1)+
      geom_line(aes(y=fit, col="pred"), lwd=1)+
      geom_vline(xintercept = as.numeric(cutoff_date), linetype = 4) +
      ylab(dep.vars[i]) +
      theme_classic() +
      labs(col = "") +
      ggtitle("")
  }

(
  plot.reg(1)
  / plot.reg(2)
  / plot.reg(3)
)
```



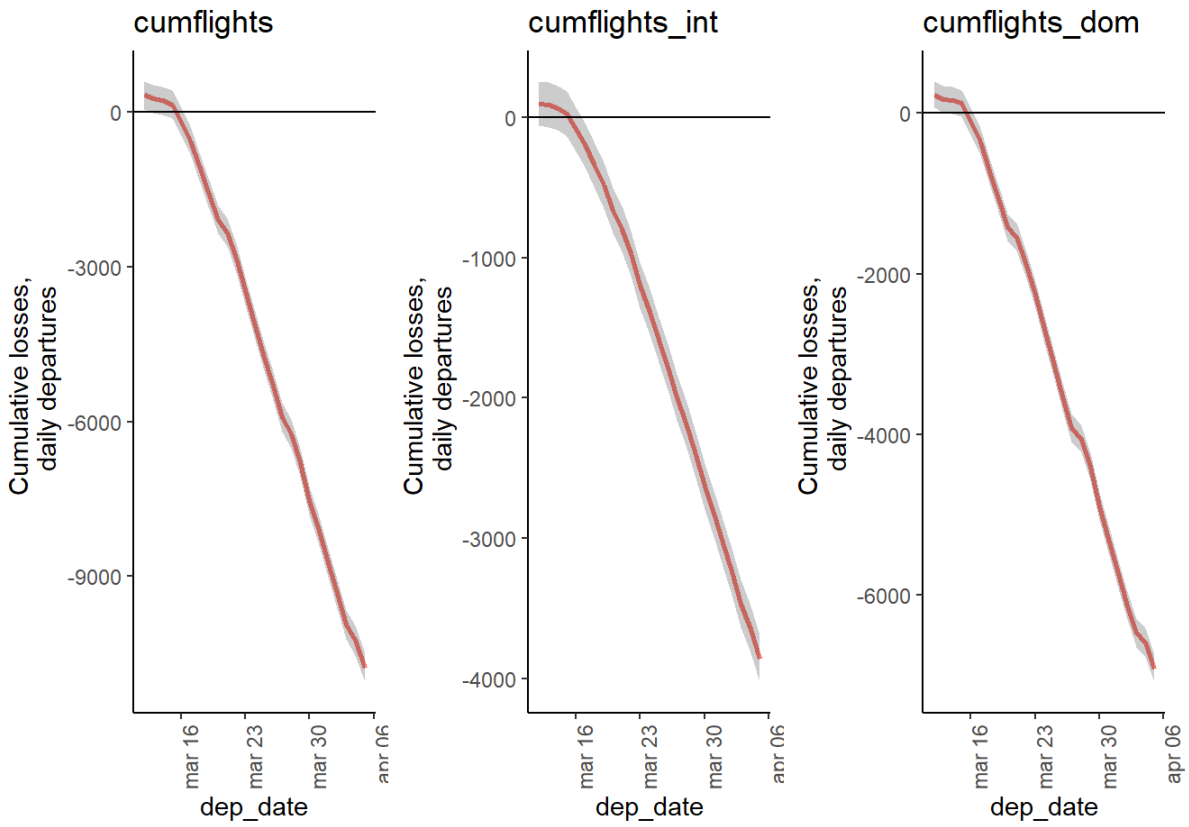

Similarly, we can see the the cumulative daily flights have a markedly lower growth rate after the crisis.

```
(
  plot.reg(4)
  / plot.reg(5)
  / plot.reg(6)
)
```



To quantify the effect of the crisis, we can instead plot the *difference* between the expectation and the realizations for the cumulative daily flights. Further, we can include 95% prediction intervals. We can note that all the time series start out slightly above zero. This might seem odd, and means that the cumulative number of flights were slightly higher than what the model predicted at the cutoff time. We could perhaps adjust this plot to force it to start at zero at the cutoff date, but it is also possible that there were indeed more flights right around the cutoff time due to the corona crisis (if, e.g. some flights were undertaken in expectation of coming travel limitations). However, such an adjustment makes little difference to the overall effect. At the time of writing (March 20th), the cumulative loss in departures were around 2113 flights $\pm$ 273.5 for a 95% prediction interval, so the effect of not starting at zero is negligible.

```
plot.losses <-
  function(i){
    df_total %>%
      bind_cols(predictions[i]) %>%
      filter(dep_date >= cutoff_date) %>%
      mutate(
        lwr = (get(dep.vars[i])-lwr),
        upr = (get(dep.vars[i])-upr),
        fit = (get(dep.vars[i])-fit)) %>%
      ggplot(aes(x=dep_date))+
      geom_line(aes(y=fit, col="pred"), lwd=1)+
      geom_ribbon(aes(ymin=lwr,ymax=upr), alpha=.25)+
      geom_hline(yintercept = 0) +
      ggtitle(dep.vars[i]) +
      theme_classic() +
      ylab("Cumulative losses, \n daily departures")+
      theme(legend.position = "none") +
      theme(axis.text.x = element_text(angle = 90, hjust = 1))
  }

plot.losses(4) + plot.losses(5) + plot.losses(6)
```

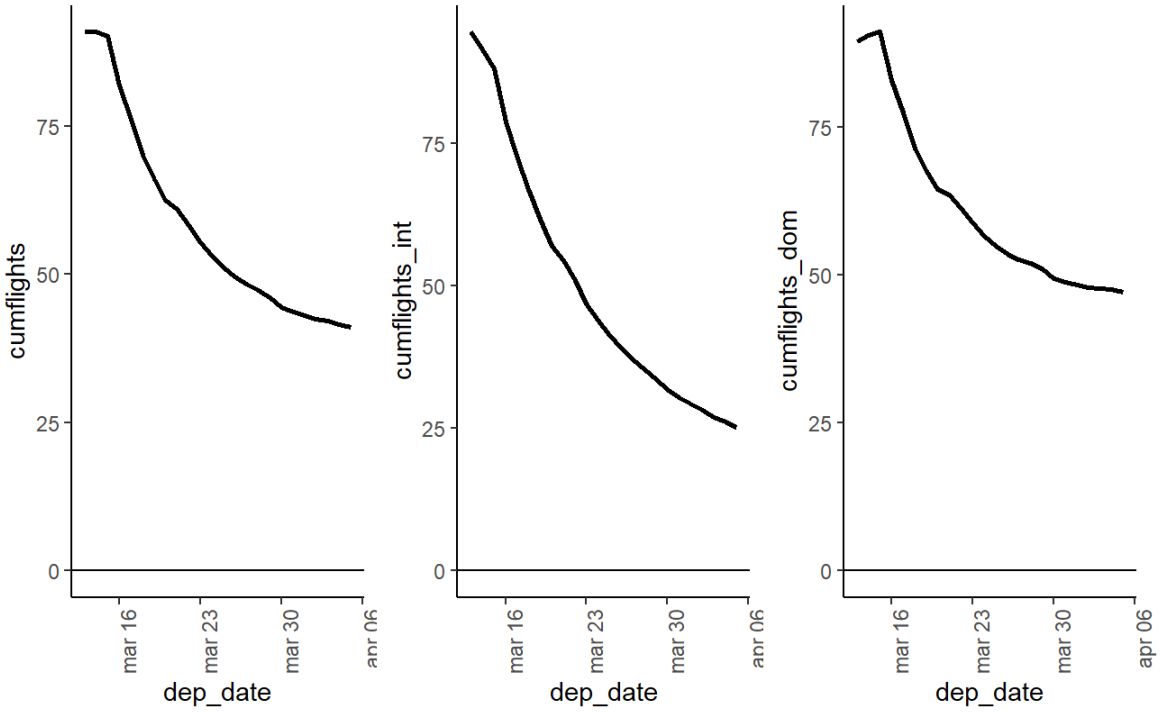



An alternative exposition is to display traffic departures as an fraction of the expection from the pre-crisis period. The only trick here is to be careful with which time period we are normalizing to. If we use the raw fraction of cumulative flights divided by predicted cumulative flights, we will hardly detect any effect at all as a few thousand missing flights is a small percentage of all flights in 2020. However, if we measure cumulative post-crisis flights relative to cumulative predicted post-

crisis flights, we only observe the share of missing flights after the crisis emerged. Hence, this graph is a nice complement to the figures over missing flights in absolute terms. At the time of writing, the total number of flights is around 40% lower than what we would expect from pre-crisis traffic patterns.

```
plot.losses.percent <-
  function(i){
    df_total %>%
      bind_cols(predictions[i]) %>%
      mutate(
        dep_var = as.numeric(!!sym(dep.vars[i])),
        dep_var_at_cutoff =
          max(
            case_when(
              dep_date == cutoff_date~dep_var,
              TRUE~0),
            na.rm=T),
        fit_at_cutoff =
          max(
            case_when(
              dep_date == cutoff_date~fit,
              TRUE~0),
            na.rm=T)) %>%
      mutate(
        fit = (
          100
          * (dep_var - dep_var_at_cutoff)
          / (fit - fit_at_cutoff)
          - 1)
        ) %>%
      filter(dep_date >  cutoff_date) %>%
      ggplot(aes(x = dep_date)) +
      geom_line(aes(y = fit), lwd = 1) +
      geom_hline(yintercept = 0) +
      ylab(paste0(dep.vars[i])) +
      theme_classic() +
      ggtitle("") +
      theme(legend.position = "none") +
      theme(axis.text.x = element_text(angle = 90, hjust = 1))
  }

p <-
  (
    plot.losses.percent(4)
  + plot.losses.percent(5)
  + plot.losses.percent(6)
  )

wrap_elements(p) +
  ggtitle("Traffic departures as percent of expectation, after Corona crisis ")
```

Traffic departures as percent of expectation, after Corona crisis




There are several ways we could improve the model estimates. An incomplete list of possible improvements is:

- Use a longer estimation period, so we can account for low frequency seasonalities, and hence, obtain a more credible long-term estimate of the effect of the crisis.
- Use a more advanced time series model, to account for the dependencies in the residuals.

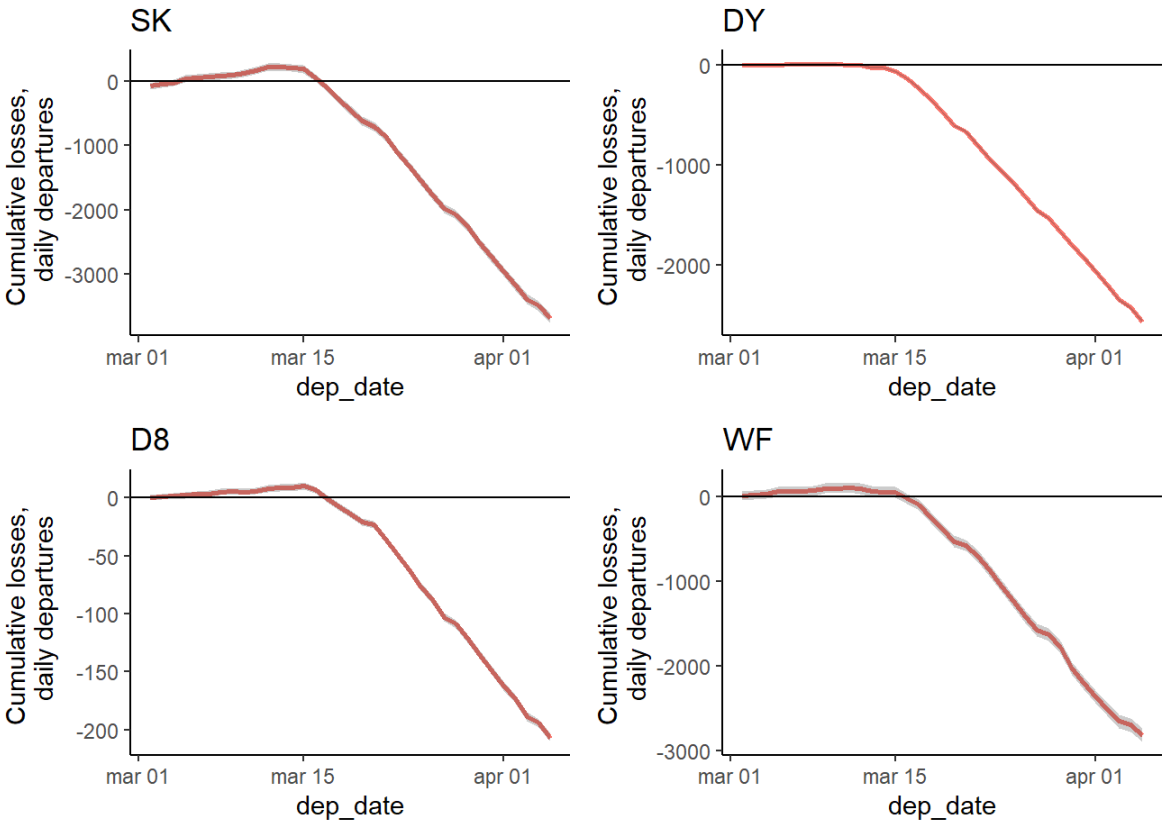
However, although the model is simple and rough, I think it gives a reasonable estimate of the magnitude of the crisis in the Norwegian airline industry.

# Q4

This question is very open, and hence there isn't much scope for writing a solution. See the rubric for home exam assessment for pointers on how questions are graded.

As a brief example though, we can repeat the analysis above for individual airlines. Below are plots of cumulative losses for SAS, Norwegian (DY), the international part of Norwegian (D8) and Widerøe. At the time of writing (March 20th), cumulative losses for SAS and Norwegian are close to 500 flights, and almost 400 for Widerøe. It will be interesting, to say the least, to see what long term consequences the crisis has on the airline industry - both in Norway in globally.

```
airlines <- c("SK","DY","D8","WF")
regressions <- vector("list", length(airlines))
predictions <- vector("list", length(airlines))

for(i in 1:length(airlines)){
  regressions[[i]] <-
    lm(
      cumflights~trend + day_of_week,
      data=df_airline,
      subset = (dep_date<cutoff_date & airline==airlines[i]))

  predictions[[i]] <-
    predict(
      regressions[[i]],
      newdata = df_airline %>% filter(airline==airlines[i]),
      interval = "confidence") %>%
    as.data.frame()
}


plot.losses <-
  function(i){
    df_airline %>%
      filter(airline==airlines[i]) %>%
      bind_cols(predictions[i]) %>%
      filter(dep_date > as.Date('2020-03-01')) %>%
      mutate(
        lwr = (cumflights - lwr ),
        upr = (cumflights - upr ),
        fit = (cumflights - fit )) %>%
      ggplot(aes(x=dep_date))+
      geom_line(aes(y=fit, col="pred"), lwd=1)+
      geom_ribbon(aes(ymin=lwr,ymax=upr), alpha=.25)+
      geom_hline(yintercept = 0) +
      theme_classic() +
      ggtitle(airlines[i])+
      ylab("Cumulative losses, \n daily departures")+
      theme(legend.position = "none")
  }

plots <- lapply(1:length(airlines), plot.losses)

patchwork::wrap_plots(plots)
```


SK
DY
D8
WF
Cumulative losses, daily departures
dep_date
mar 01
mar 15
apr 01
0
-1000
-2000
-3000
-50
-100
-150
-200