Question5

Yufei Yin

Question 5

Option B: Rolling Linear Regression for Local Weather (10 marks)

This question pertains to predicting the weather locally and on a small timescale. We will investigate two ways of predicting the next day's average temperature in Vancouver based on historical data. The two ways are as follows:

1) A linear interpolation (linear regression) based on the previous three days, extrapolating to the next day (the day right after those three days).
2) A prediction that the next day's average temperature in Vancouver is exactly the same as the average temperature in Vancouver of the immediately preceding day. We will score these predictions based on root mean squared error (RMSE). This is the L-2 norm between the predicted and actual average daily temperature values divided by the number of values (if y is a vector, and y0 is a prediction of y then the RMSE between y and y0 is sqrt(mean((y-y0)^2)) in R code—note that y and y0 must have the same length).

a)

Download daily climate data (Climate Daily/Forecast/Sun) from https://vancouver.weatherstats.ca/download.html including the 7-day period between Monday the 12th of April 2021 and Sunday 18th of April 2021 (inclusive). Extract this 7-day period (and provide it in a listing) and report the mean average hourly temperature (avg_hourly_temperature) for all 7 days, and the standard deviation of the average hourly temperature for all 7 days. (4 marks)

```r
data = read.csv('weatherstats_vancouver_daily.csv')
data = data[10:16,]
data = data[order(data$date),]
data$date
```

```
## [1] "2021-04-12" "2021-04-13" "2021-04-14" "2021-04-15" "2021-04-16"
## [6] "2021-04-17" "2021-04-18"
```

```r
# mean average hourly temperature
mean(data$avg_hourly_temperature)
```

```
## [1] 10.80143
```

```r
# standard deviation of the average hourly temperature
sd(data$avg_hourly_temperature)
```

```
## [1] 2.134225
```

b)

For each day between Thursday 15th of April 2021 and Sunday 18th of April 2021 (inclusive), fit a linear regression model (using the R function lm for example) with xvalues “1, 2, 3” and y-values given by the avg_hourly_temperature for the three days immediately preceding that day. Then, predict the avg_hourly_temperature for that day by extrapolating the linear regression for the x-value “4”. What are the predictions for the 4 days under question? And what is the RMSE between the predictions and the actual avg_hourly_temperature values? (2 marks)

```
days = 1:3
# April 15
Apr15_mdl = lm(formula = avg_hourly_temperature ~ days, data = data[1:3,])
Apr15_pred = predict(Apr15_mdl, newdata = data.frame(days = 4))
Apr15_pred
```

```
##        1
## 10.80667
```

```
# April 16
Apr16_mdl = lm(formula = avg_hourly_temperature ~ days, data = data[2:4,])
Apr16_pred = predict(Apr16_mdl, newdata = data.frame(days = 4))
Apr16_pred
```

```
##        1
## 12.36333
```

```
# April 17
Apr17_mdl = lm(formula = avg_hourly_temperature ~ days, data = data[3:5,])
Apr17_pred = predict(Apr17_mdl, newdata = data.frame(days = 4))
Apr17_pred
```

```
##        1
## 13.83667
```

```
# April 18
Apr18_mdl = lm(formula = avg_hourly_temperature ~ days, data = data[4:6,])
Apr18_pred = predict(Apr18_mdl, newdata = data.frame(days = 4))
Apr18_pred
```

```
##     1
## 13.28
```

```
# RMSE
y = data[4:7,]$avg_hourly_temperature
y0 = c(Apr15_pred, Apr16_pred, Apr17_pred, Apr18_pred)
sqrt(mean((y-y0)^2))
```

```
## [1] 0.7068946
```

c)

Consider again the 4 days between Thursday 15th of April 2021 and Sunday 18th of April 2021 (inclusive). What if we predict the avg_hourly_temperature for each day using the avg_hourly_temperature of the previous day (i.e., predict tomorrow's avg_hourly_temperature with today's avg_hourly_temperature)? Compute and report the RMSE between these predictions and the actual avg_hourly_temperature values. This is the martingale assumption (this assumption is often applied in finance). (2 marks)

```r
y = data[4:7,]$avg_hourly_temperature
y0 = data[3:6,]$avg_hourly_temperature
# predictions
y0
```

```
## [1]  9.41 11.61 12.18 12.72
```

```r
# RMSE
sqrt(mean((y-y0)^2))
```

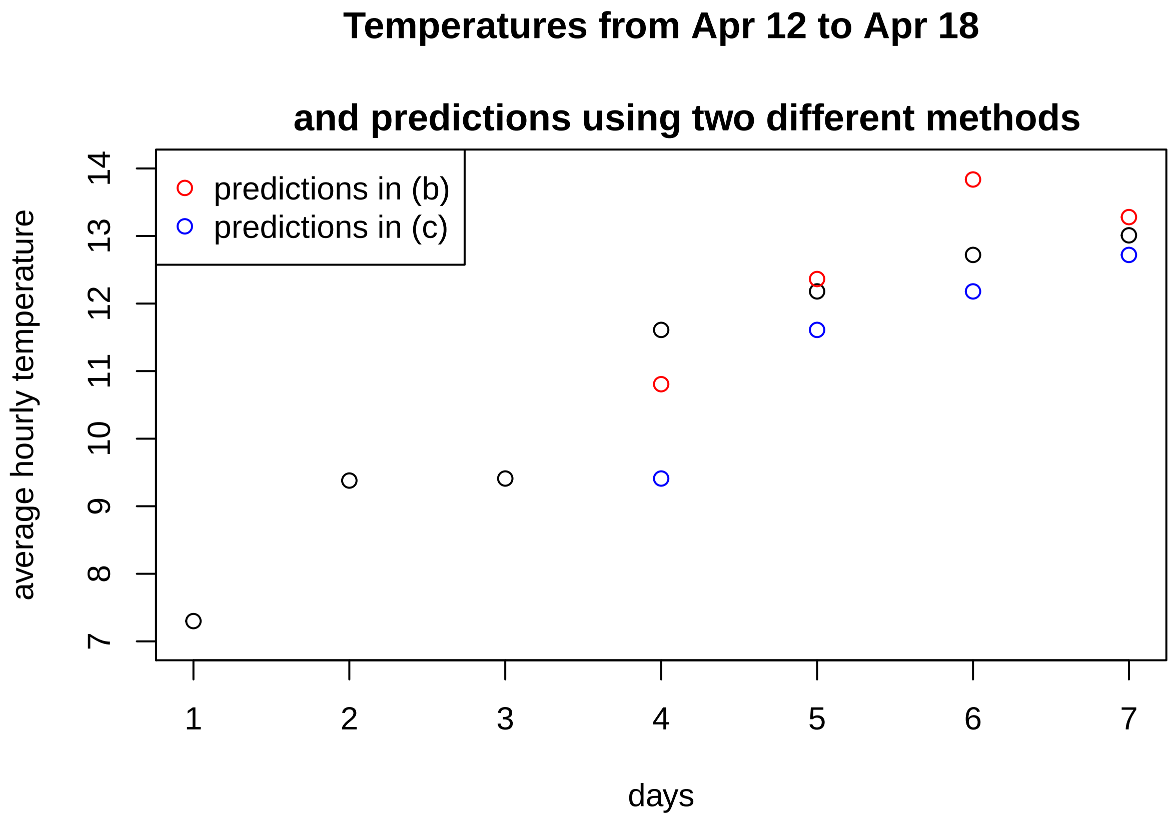
```
## [1] 1.176924
```

d)

What's better for predicting weather, according to this analysis: The predictions using the three-day rolling linear regression (b), or the predictions using the previous day's value (c)? Provide a reason as to why this may be the case. (2 marks)

According to the analysis, the predictions using the three-day rolling linear regression is better for predicting weather. RMSE in (b) is closer to 0 compared to RMSE in (c) which indicates the predictions are closer to actual values in (b).

```
plot(1:7, data$avg_hourly_temperature,
     xlab = 'days',
     ylab = 'average hourly temperature',
     ylim = c(7,14),
     main = "Temperatures from Apr 12 to Apr 18\n
     and predictions using two different methods")
df = data.frame(days = c(4:7),
                pred = c(Apr15_pred, Apr16_pred, Apr17_pred, Apr18_pred))
points(df$days, df$pred, col = "red")
df2 = data.frame(days = c(4:7),
                 pred = data[3:6,]$avg_hourly_temperature)
points(df2$days, df2$pred, col = "blue")
legend(x = "topleft", col = c("red","blue"), pch = c(1,1),
       c("predictions in (b)", "predictions in (c)"))
```



According to the plot above, we noticed that from day3(April 14) to day4(April 15), there is a significant increase in average hourly temperature. If we use the martingale assumption to predict April 15's average hourly temperature, the prediction in April 15 is far away from the actual average hourly temperature in April 15 (black point and blue point in day 4); moreover, it also causes the prediction in April 17 using three-day rolling linear regression not close to the actual value in April 17(black point and red point in day 6). Overall, the predictions using three-day rolling linear regression (red points) are closer to the actual values (black points) which also can be proof by the RMSE values we calculated before.