Homework 2 – Exercise 1.8 Question 1

1 For case 3, a series of predictor variables can be considered to forecast the resale values of vehicles:

- Time trend. This is to capture the trajectory of the resale value over time and capture the effects of relevant variables in the regression model that change over time and not directly measurable (eg, technological change). There might be an upward or downward trend on the resale price as demand for resale cars might fluctuate over different period of time (eg, business cycles)

- Seasonal dummy variables. This is to include monthly seasonality as demand for resale cars might fluctuate in a deterministic pattern across different months. For example, resale price value may increase periods at the end & start of each year as this period coincides with individuals getting their employment bonuses.

- Spike variables. We can also include events/effects that last for only one period (unlike seasonality) such as industrial action and rise of pandemic, etc. This is equivalent to a dummy variable for handling an outlier which we might expect a significant change in resale price.

- Distributed lags. Asides from including advertising expenditure (for example) for the time period, we can also include lagged values of advertising expenditure since the effect of advertising can last beyond the actual campaign.

- Attributes of vehicles. We can include the different type of vehicles (cars, motorcycles, vans, etc), colour of vehicles (blue, red, etc), horsepower, auto/manual, etc corresponding to its resale value to better forecast for each type and characteristic of vehicle.

- Customer profile. We can include the demographics such as the age group, income level, household size, etc, of the customers. The better understanding of the target audience for each type of vehicle allows for better forecast of demand for the cars and its corresponding resale price.

For case 4, a series of predictor variable can be considered to forecast weekly air passenger traffic:

- Time trend, seasonal dummy variables (school holidays), spike variables (pilot strike, major sporting events, flight promotions, flight-related news such as airplane crash over the past month, rise of pandemic which leads to travel bans), distributed lags as explained in Case 3

- Step Variable & piecewise linear trend. We can include a step variable to account for the effects of the new air-line cut-price and the seat reclassification

- Flight characteristics. We would include many characteristics of the flight such as duration of flight, destination & its immigration requirements (visa required), type of travel (economic/business/first-class), average boarding & landing time which may affect customers' preference for the air carrier, etc

- Flight prices & Oil price. As the frequency of air passengers is highly dependent on the flight prices and flight prices are correlated with oil prices, it is useful to include oil prices as the predictor variable.

- Macroeconomic factors such as exchange rate and unemployment rate that might affect the purchasing power of the customers.

Homework 2 – Exercise 5.1 Question 1

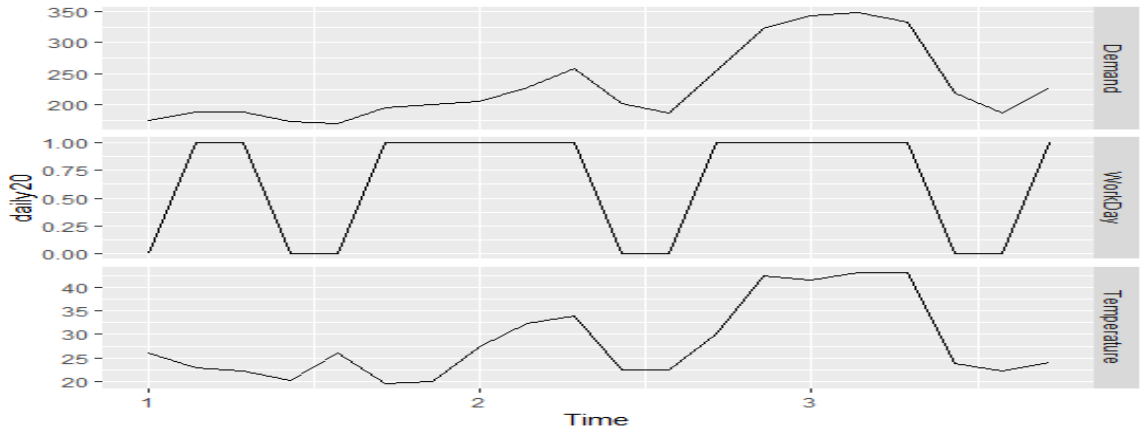
a) Plot the data and find the regression model with temperature as an explanatory variable. Why is there a positive relationship?

```
library(fpp2)
daily20 <- head(elecdaily,20)
#Plot the data and find the regression model with temperature as an explanatory variable
autoplot(daily20[,c(1, 3)], facets=TRUE)
autoplot(daily20, facets=TRUE)
daily20 %>%
  as.data.frame() %>%
  ggplot(aes(x=Temperature, y=Demand)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```

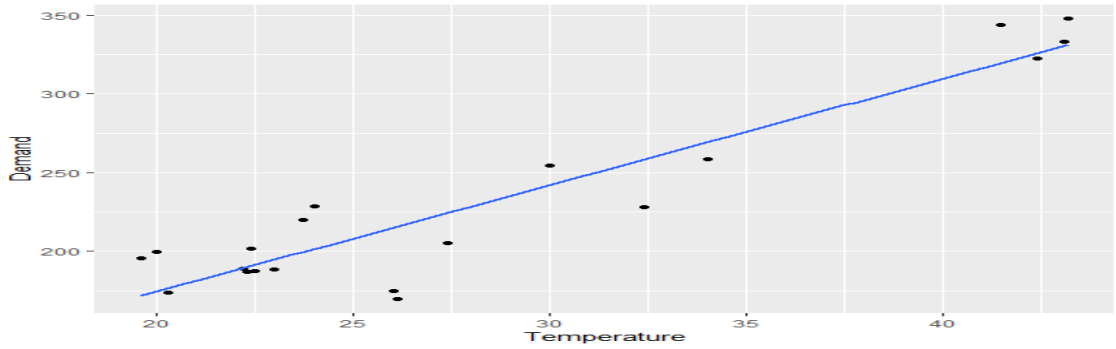
A quick look at data20 object

	Demand	WorkDay	Temperature
1	174.8963	0	26.0
2	188.5909	1	23.0
3	188.9169	1	22.2
4	173.8142	0	20.3
5	169.5152	0	26.1
6	195.7288	1	19.6
7	199.9029	1	20.0
8	205.3375	1	27.4
9	228.0782	1	32.4
10	258.5984	1	34.0
11	201.7970	0	22.4
12	187.6298	0	22.5
13	254.6636	1	30.0
14	322.2323	1	42.4
15	343.9934	1	41.5
16	347.6376	1	43.2
17	332.9455	1	43.1
18	219.7517	0	23.7
19	186.9816	0	22.3
20	228.4876	1	24.0

Plot the data using autoplot



Plot the scatterplot with demand and temperature



Find the regression model

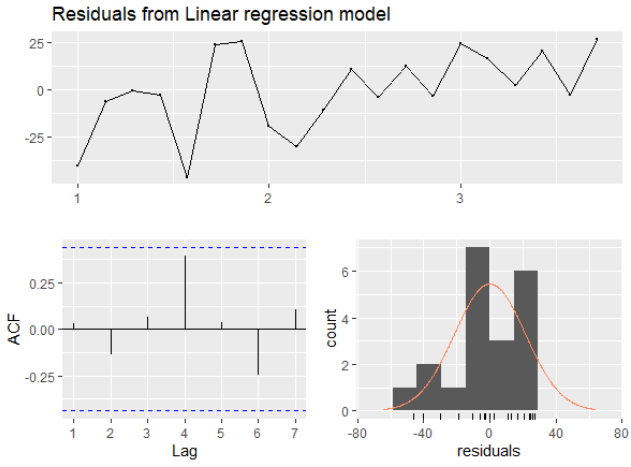
```
15  fit <- tslm(Demand ~ Temperature, data=daily20)
16  summary(fit)
```

```
Call:
tslm(formula = Demand ~ Temperature, data = daily20)

Residuals:
    Min      1Q  Median      3Q     Max
-46.060  -7.117  -1.437  17.484  27.102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.2117    17.9915   2.179   0.0428 *
Temperature   6.7572     0.6114  11.052 1.88e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22 on 18 degrees of freedom
Multiple R-squared:  0.8716,    Adjusted R-squared:  0.8644
F-statistic: 122.1 on 1 and 18 DF,  p-value: 1.876e-09
```

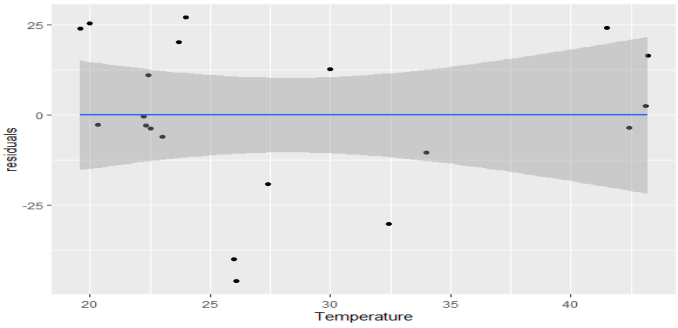
From the scatterplot and the summary, I observe that temperature is statistically significant in explaining the variation in electricity demand and there is a positive relationship between them. This is sensible as electrical appliances such as fans and air-conditioner will be more heavily used during periods of high temperature to cool the environment. Hence, demand for electricity increases as temperature increases.

b) Produce a residual plot. Is the model adequate? Are there any outliers?

```
17  #Produce a residual plot, is the model adequate? Are there any outliers?
18  checkresiduals(fit)
19  df <- as.data.frame(daily20)
20  df[,"residuals"] <- as.numeric(residuals(fit))
21  ggplot(df, aes(x=Temperature, y=residuals)) + geom_point() + geom_smooth(method="lm")
```



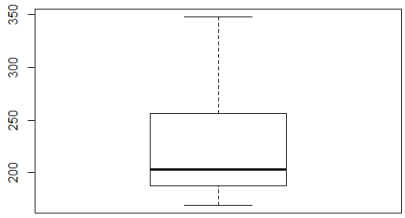
LM test shows a p-value = 0.5774



The ACF plot & results from LM test shows that the residuals are not autocorrelated with its lags (up till 7 lags) at 5% significance level (p=0.5774 > 0.05). Also, the residuals do not seem to be correlated with Temperature. Hence, the model is adequate for forecasting. However, the histogram shows that the residuals are not normally distributed, and this may cause the prediction interval to be inaccurate. The distribution of the residuals could be estimated by using a bootstrapping method for better prediction interval estimation or the model could be improved to yield better forecasts.

Boxplot to show outliers

```
> boxplot(daily20[, "Demand"])
```



It is difficult to detect for outliers with a small sample size of 20. From the boxplot, it shows that all the data lies within the inter-quantile range. However, from the residuals timeplot and histogram, it might show evidence that there are outliers. I would investigate observations 14, 15 & 16 as these observations show periods of high electricity demand (>300) corresponding to high temperatures (>40) to be more conservative.

c) Use the model to forecast the electricity demand that you would expect for the next day if the maximum temperature was 15 and compare it with the forecast if the maximum temperature was 35. Do you believe these forecasts?

```
29 #Scenario-based forecasting
30 newdata <- data.frame(Temperature = c(15,35))
31 predicted <- forecast(fit, newdata=newdata)
32 predicted
```

```
         Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
3.857143       140.5701 108.6810 172.4591  90.21166 190.9285
4.000000       275.7146 245.2278 306.2014 227.57056 323.8586
```

When the temperature was 15 degrees, the point forecast for electricity demand was 140.6. When the temperature was 35 degrees, the point forecast for electricity demand was 275.7. I find the forecast for both scenarios to be reasonable as they were near to the range of temperatures in the data. Since this is a one-day ahead forecast

However, I find it easier to believe the forecast for temperature = 35degrees as compared to temperature = 15degrees as the former is an interpolation (and the prediction is similar to the actual data when temperature = 34degrees) whereas the latter is an extrapolation to the results that might yield misleading results as we have no information on temperatures below 19.6 degrees. Logically, I would expect the demand for electricity when temp = 15degrees to be higher, if not the same, as compared to temp=19.6degrees since individuals will be more likely to turn on heaters at such low temperature. However, the point forecast for temp=15degrees (140.6) is significantly lower than that of when temperature = 19.6degrees (195.7 as observed in the table in part a).

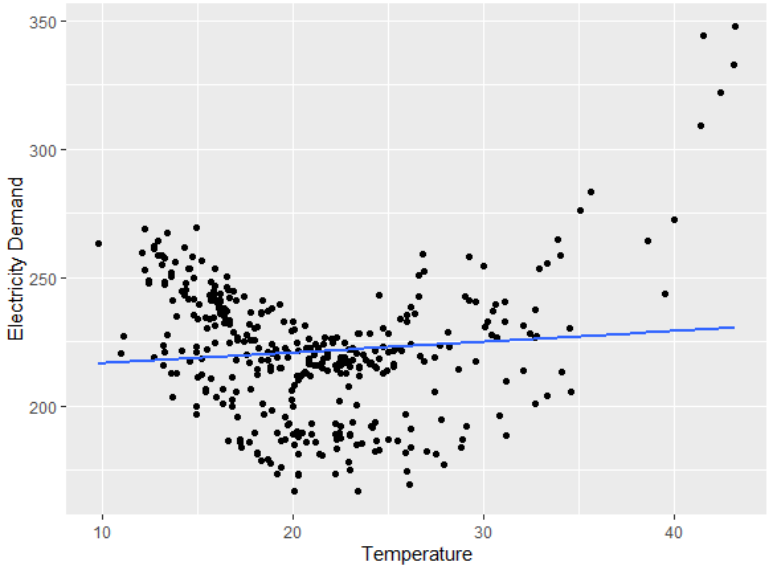
d) Give prediction intervals for your forecasts

```
         Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
3.857143       140.5701 108.6810 172.4591  90.21166 190.9285
4.000000       275.7146 245.2278 306.2014 227.57056 323.8586
```

For 15 degrees scenario, the 80% prediction interval, assuming normality, is (108.9, 172.5) and 95% interval is (90.2, 190.9).
For 35 degrees scenario, the 80% prediction interval, assuming normality, is (245.2, 306.2) and 95% interval is (227.6, 323.9).

e) Plot demand vs Temperature for all of the available data in elecdaily. What does this say about your model?

```
38 #Repeat using all of available data in elecdaily
39 elecdaily %>%
40   as.data.frame() %>%
41   ggplot(aes(x=Temperature, y=Demand)) +
42   ylab("Electricity Demand") +
43   xlab("Temperature") +
44   geom_point() +
45   geom_smooth(method="lm", se=FALSE)
```



The result plot shows that the model as described above was based on a small sample size as compared to the total data. Hence, the model could have explained the first 20 days well but the model was not right for the total data. A non-linear trend can be observed by studying the scatterplot of temperature and electricity demand with electricity demand decreasing at low temperatures and start to increase again after about 22 degrees.