Applied Time Series Analysis for Forecasting

Time series regression models

AAA Washington Case

Based on your excellent performance and great recommendations from Mr. Tux Michael DeCoria had offered you a contract as a consultant to AAA. Your goal is to assist Mr DeCoria in finding a way to predict emergency road service call volume for future years. Their previous analysis addressed the effect of average daily temperature on emergency road service call volume. It has been found that the temperature's effect is significant and could explain about half of calls volume variability. You have also used the Box-Jenkins models and found that you can predict the future AAA calls volume based on the past records with average error rate about 6%.

After discussing your findings with Michael you decide to combine the two approached and explore the ADL models.

A conversation with the manager of the emergency road service center has led to two important observations: (1) Automakers seem to design cars to operate best at 60 degrees Fahrenheit and (2) call volume seems to increase more sharply when the average temperature drops a few degrees from an average temperature in the 30s than it does when similar drop occurs with an average in the 60s. This information suggests that the effect of temperature on emergency road service is nonlinear.

Michael DeCoria stated that he believes the number of road service calls received is related to the general economic cycle and that the Washington State unemployment rate is a good surrogate measurement for the general state of Washington's economy. Now he has observed that the cyclical trend of the time series seems to be lagging behind the general economic cycle.

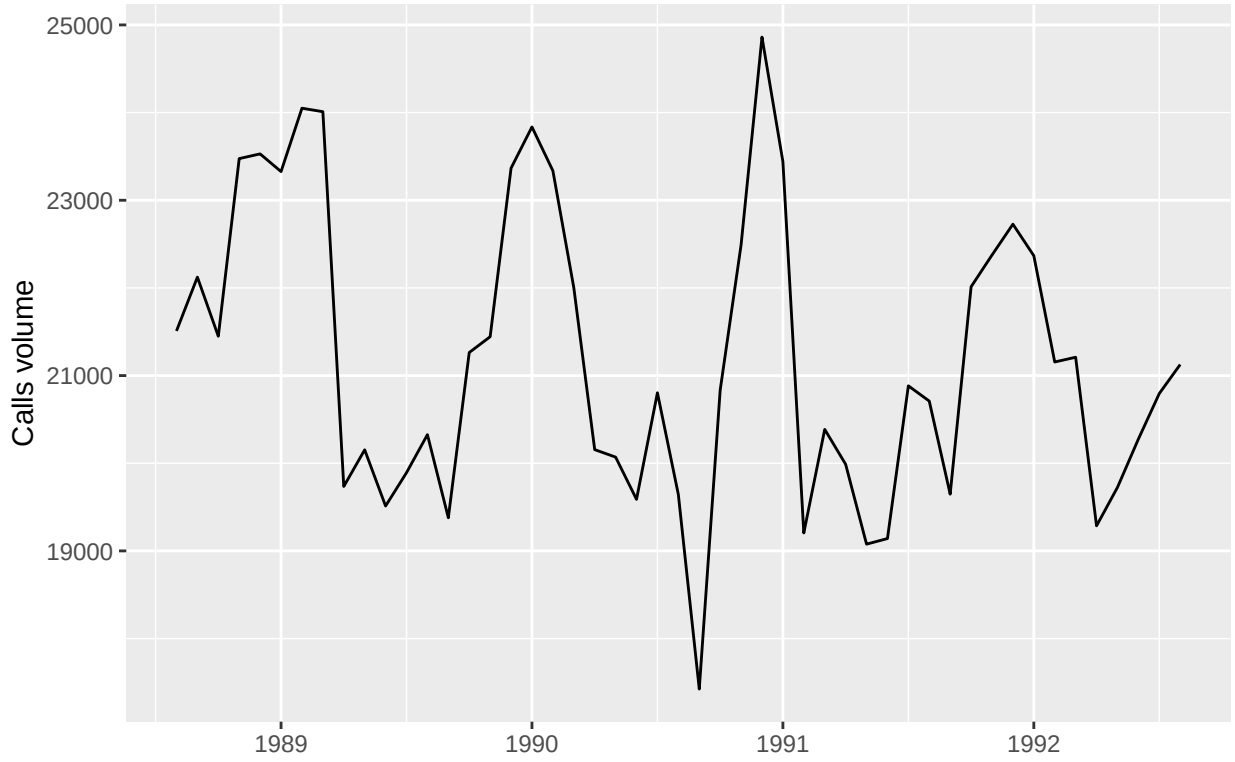
The data on emergency road service call volume, average monthly temperature and the Washington State unemployment rate are given in **AAAdatacleaned.csv**.

(1) Upload the data into R. Create time series objects for the calls, temperature, and unemployment rate variables. Plot all three variables, label the axis. Comment on the time series. Include your script.

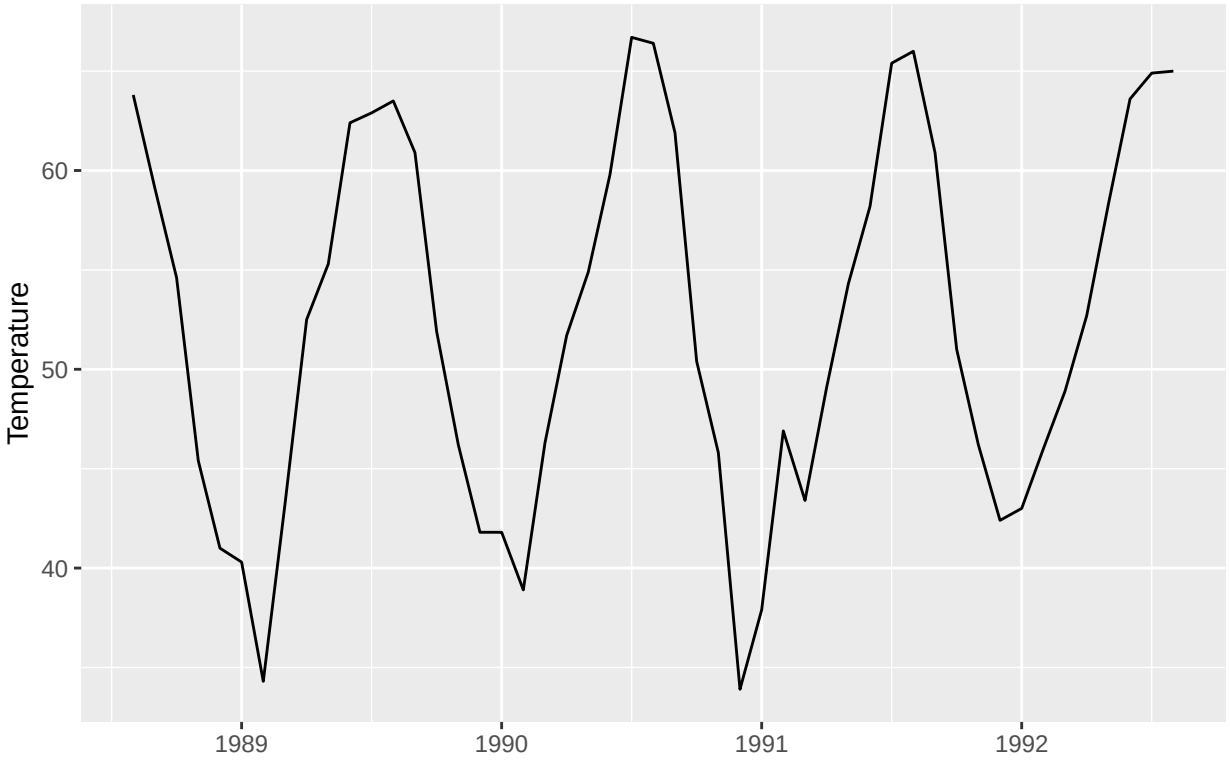
```
data = read.csv("AAAdatacleaned.csv")
Call.ts = ts(data$Calls, start = c(1988,8),
             frequency = 12)
Temp.ts = ts(data$Temp, start = c(1988,8), frequency = 12)
Rate.ts = ts(data$Rate, start = c(1988,8), frequency = 12)

library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

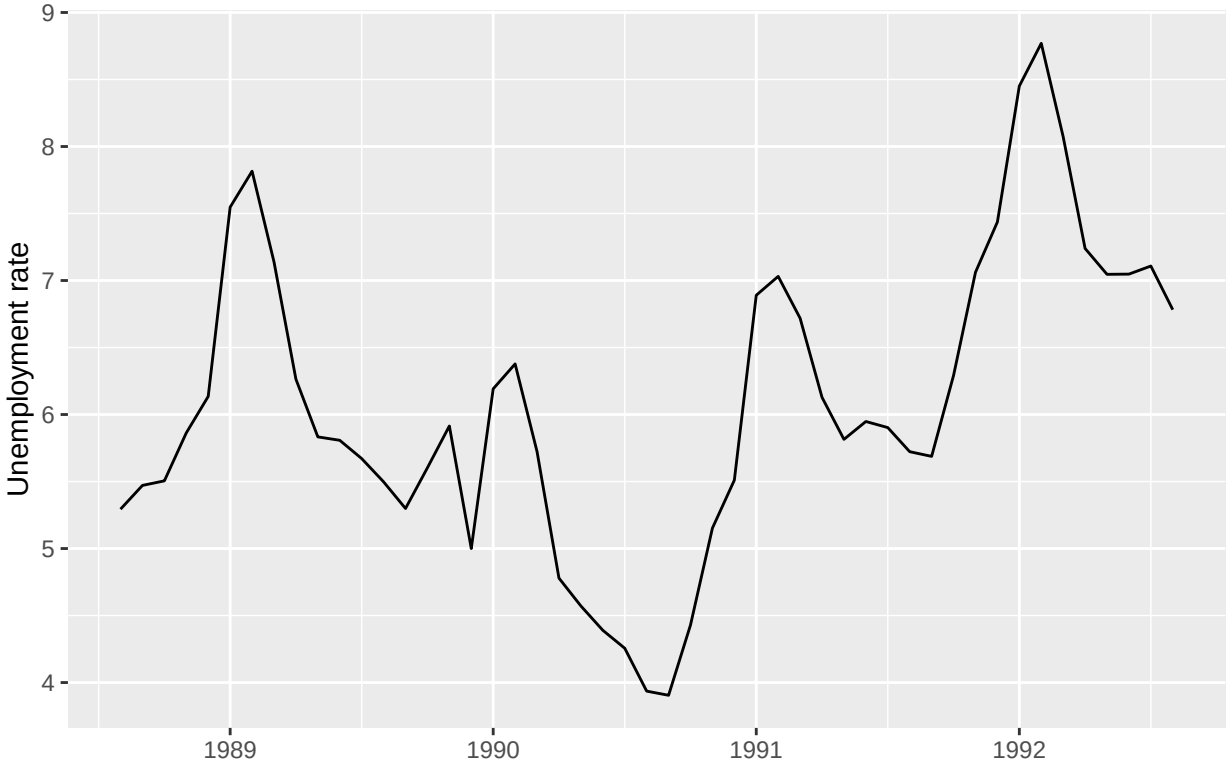
```
autoplot(Call.ts, xlab='', ylab='Calls volume')
```




```
autoplot(Temp.ts, xlab='', ylab='Temperature')
```



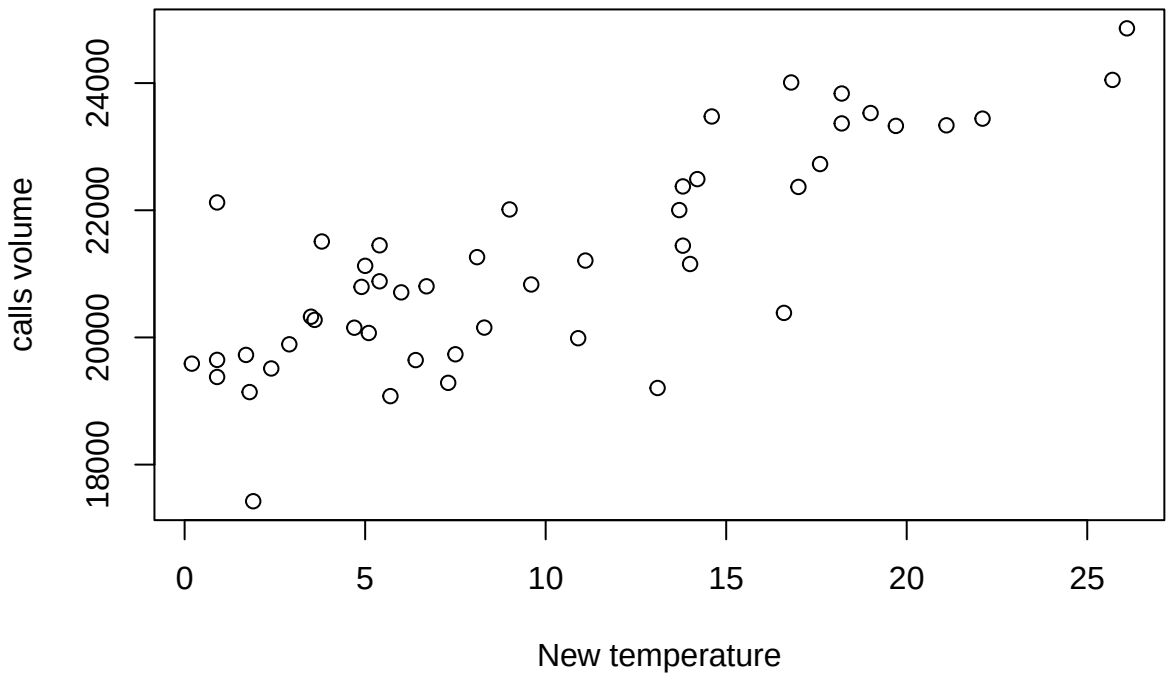

```
autoplot(Rate.ts, xlab='', ylab='Unemployment rate')
```



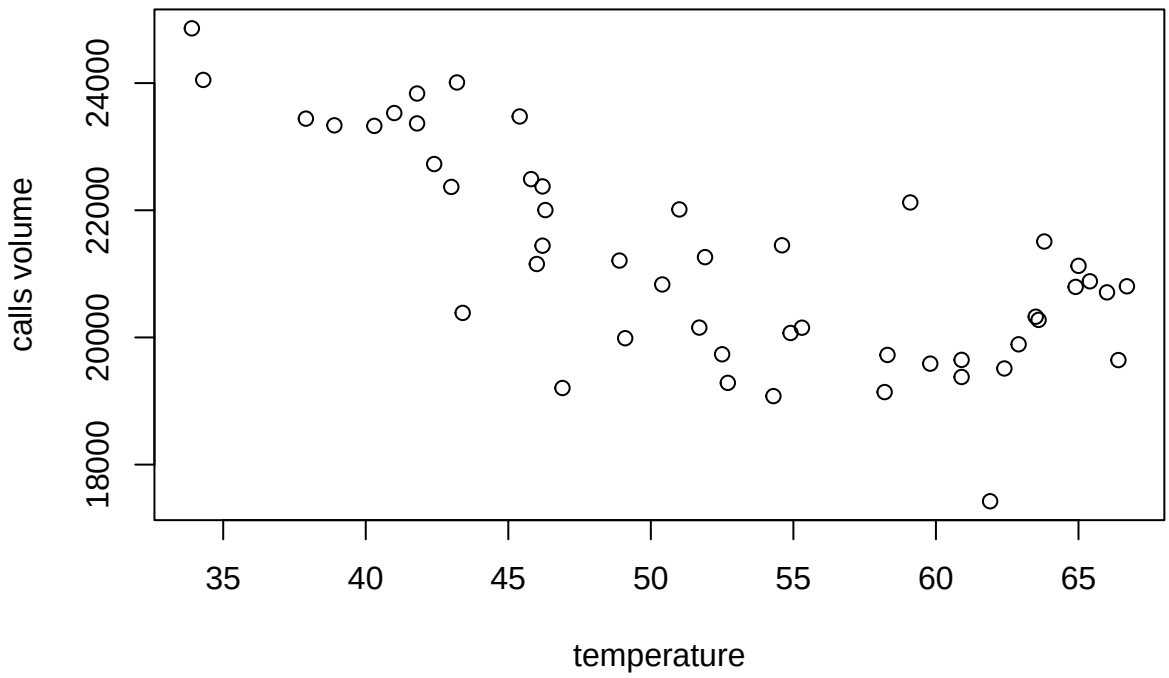
level, noise, seasonality

(2) Create a new temperature variable. Remember that temperature is a relative scale and that the selection of the zero point is arbitrary. If vehicles are designed to operate best at 60 degrees Fahrenheit, then every degree above or below 60 degrees should make vehicles operate less reliably. To accomplish a transformation of the temperature data that simulates this effect, begin by subtracting 60 from average monthly temperature values. This repositions "zero" to 60 degrees Fahrenheit. Use the absolute value of this new temperature variable. Plot two scatterplots: 1. the regular temperature vs. calls volume; 2. the new temperature variable vs. calls volume. Comment on the patterns. Include your script and the plot.

```
newtemp.ts = abs(Temp.ts - 60)
plot(as.vector(newtemp.ts), as.vector(Call.ts), xlab = "New temperature", ylab = "calls volume")
```



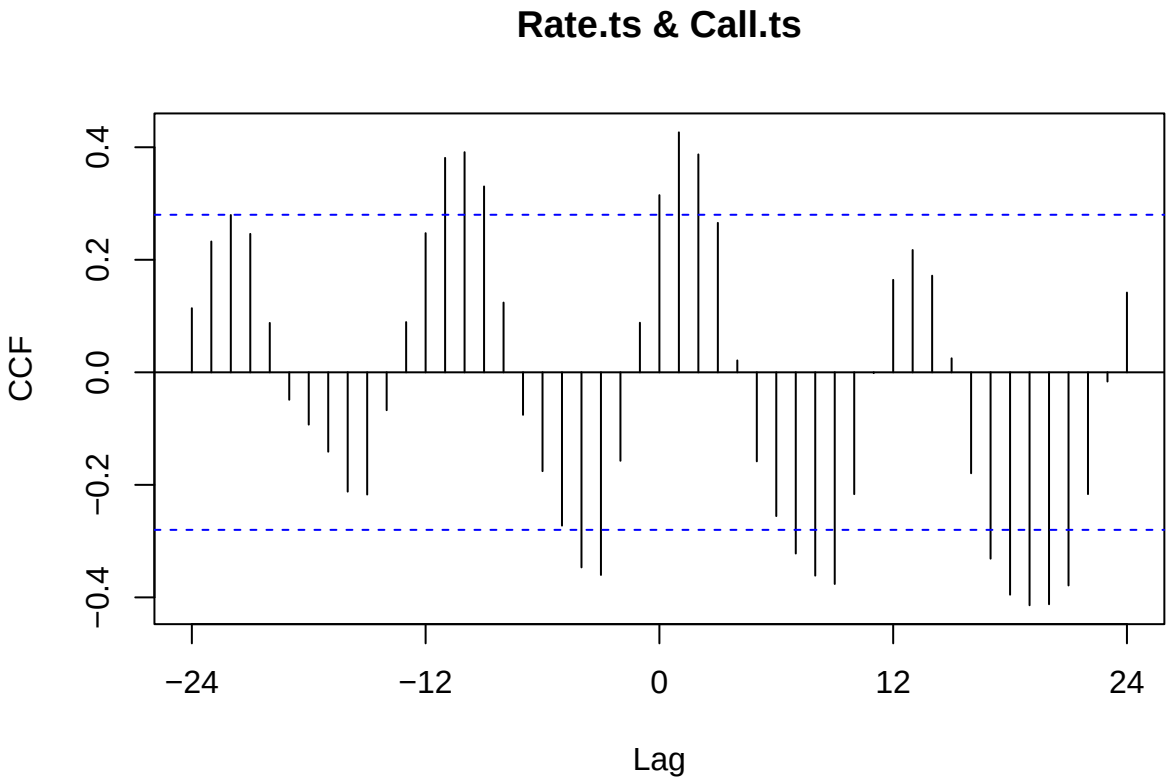

```
plot(as.vector(Temp.ts), as.vector(Call.ts), xlab = "temperature", ylab = "calls volume")
```



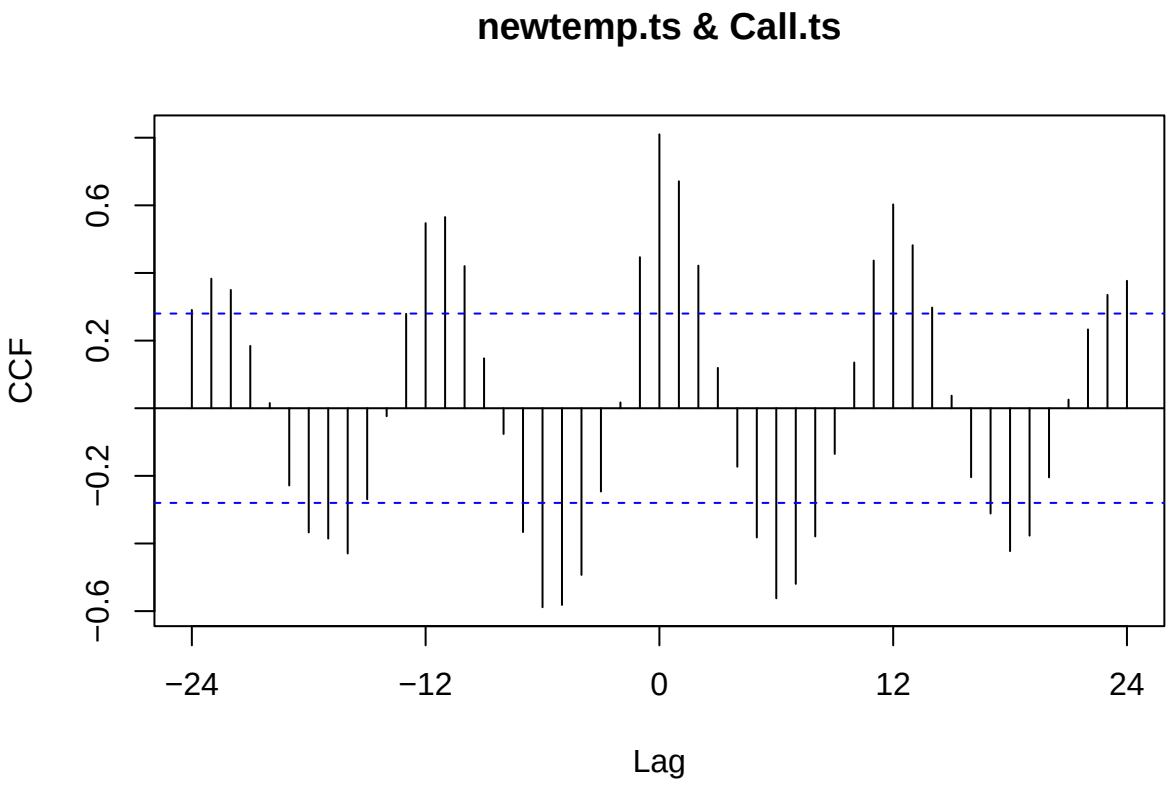
not pretty linear. the temperature below 60 seems to be downward trend, the temperature above 60 seems to be upward trend.

(3) Study the cross-correlations between the unemployment rate and the calls volume. Is the unemployment rate leading the calls and if yes, by how much? Include your script, comments, and the ccf plot.

```
Ccf(Rate.ts, Call.ts)
```




```
Ccf(newtemp.ts, Call.ts)
```



There is a significant correlation between calls volume and unemployment rates at lags -3 and -10. Let's build two models. and then choose the best one.

(4) Create a lagged unemployment rate variable and relate it to emergency road service based on your ccf analysis. Fit the model using lagged unemployment as a predictor.

Are the coefficients of the independent variables significantly different from zero? Report adjusted coefficients of determination, RMSE, and MAPE for both models.

Include your script, R output, and clear comments.

```
newdata<-ts.intersect(Call.ts, unempL3= lag(Rate.ts, -3), unempL10=lag(Rate.ts,-10), temp=newtemp.ts)
m1<-tslm(Call.ts~unempL3, data=newdata)
summary(m1)
```

```
##
## Call:
## tslm(formula = Call.ts ~ unempL3, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4256.3  -872.8    34.0   869.3  2948.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23780.0     1272.2  18.692   <2e-16 ***
```

```
## unempL3       -478.3      208.7  -2.292   0.0277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 37 degrees of freedom
## Multiple R-squared:  0.1243, Adjusted R-squared:  0.1006
## F-statistic: 5.251 on 1 and 37 DF,  p-value: 0.02772
```

```
accuracy(m1)
```

```
##                       ME    RMSE      MAE       MPE     MAPE     MASE
## Training set 3.731773e-13 1457.95 1168.193 -0.482921 5.595074 1.127318
##                    ACF1
## Training set 0.5620345
```

```
m2<-tslm(Call.ts~unempL10, data=newdata)
summary(m2)
```

```
##
## Call:
## tslm(formula = Call.ts ~ unempL10, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3687.5  -789.5   -67.6   619.2  3289.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15235.7     1358.3  11.217 1.83e-13 ***
## unempL10       993.5      234.7   4.234 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1313 on 37 degrees of freedom
## Multiple R-squared:  0.3263, Adjusted R-squared:  0.3081
## F-statistic: 17.92 on 1 and 37 DF,  p-value: 0.0001458
```

```
accuracy(m2)
```

```
##                       ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 2.798716e-13 1278.754 982.8276 -0.3711263 4.716643 0.9484379
##                    ACF1
## Training set 0.3594786
```

The second model that uses the unemployment rate from 10 months ago is better at predicting the number of the service calls. It has a positive correlation, which is expected, i.e. the higher the unemployment the higher the number of calls. Even though, the coefficients for the unemployment variables are significant for both model, model 2 is highly significant and model 1 is barely significant. Model 2 is more accurate.

(5) Develop a multiple regression equation using the new transformed average temperature variable created in and the lagged unemployment variable created earlier that describe the calls volume the best.

Is this a good model? Report the adjusted coefficients of determination, RMSE, and MAPE.

Have any of the underlying assumptions been violated?

```
m3<-tslm(Call.ts~temp + unempL10, data=newdata)
summary(m3)
```

```
##
## Call:
## tslm(formula = Call.ts ~ temp + unempL10, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2274.7  -421.5   199.0   470.5  1387.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17008.46     881.99  19.284  < 2e-16 ***
## temp          171.57      22.47   7.635 4.91e-09 ***
## unempL10      399.75     166.31   2.404   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 822.4 on 36 degrees of freedom
## Multiple R-squared:  0.7428, Adjusted R-squared:  0.7285
## F-statistic: 51.98 on 2 and 36 DF,  p-value: 2.428e-11
```

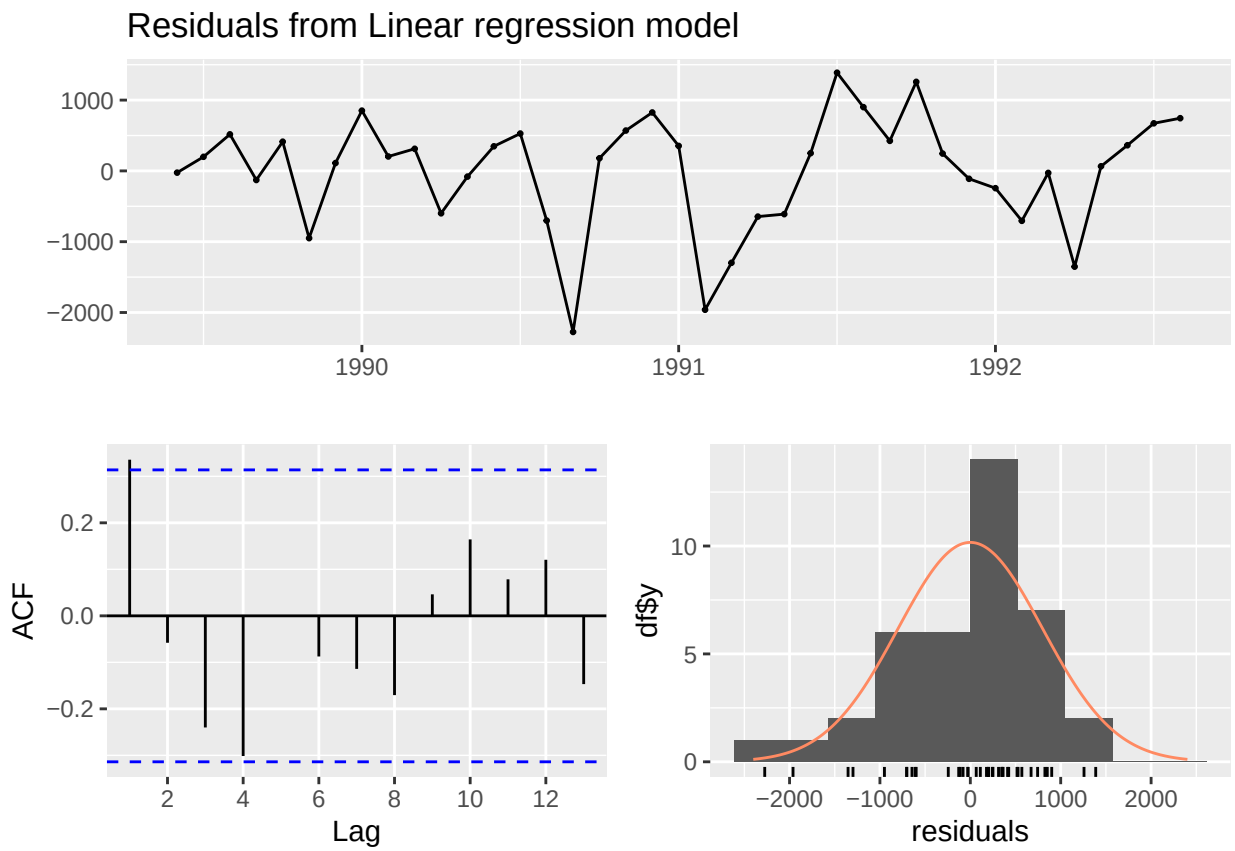
```
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: Call.ts
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## temp       1 66408144 66408144 98.1855 7.929e-12 ***
## unempL10   1  3907826  3907826  5.7778    0.0215 *
## Residuals 36 24348730   676354
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
accuracy(m3)
```

```
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 1.865716e-13 790.1433 600.8607 -0.148965 2.943097 0.5798363
##                    ACF1
## Training set 0.3357316
```

```
checkresiduals(m3)
```




```
##
##  Breusch-Godfrey test for serial correlation of order up to 8
##
## data:  Residuals from Linear regression model
## LM test = 13.714, df = 8, p-value = 0.08954
```

R-squared is not bad, but RMSE is lower than previous two models. it is fairly normal, However, it is a little not independent, see ACF at lag 1. So the model need to improve.

(6) Prepare a memo to Michael recommending the regression model you believe is more appropriate for predicting the emergency road service call volume. Write it in a paragraph form and keep in mind that Mr. DeCoria is not an expert in time series analysis or R. Therefore, don't be too technical.

Mr. DeCoria, After creating multiple models, we have found that a model that considers both temperature and unemployment is the best fit. Using the information that cars perform most reliably at 60 degrees, we made changes to the data in order to identify if there is a relationship between temperature and car breakdowns. The result was that there is a linear relationship between temperature and breakdowns, so it should be used in the model. Unemployment also proved to be significant, when looking at it multiple data points in the past, or considering the long term effect. After finding that both temperature and unemployment are significant, we created a model that combines both variables. As of now, the model we have build has an error rate of 3.13%, so you can expect it to be off by this percentage. The model meets the assumptions necessary for us to consider it acceptable. However, there are still some problems with the model, as the errors seem to have some correlation. In order to improve, we should build a model that incorporates past calls.