Multi-Variate Time Series Data for Sales Forecasting by Strategy Type

Susan Kight

Preparation

```
# install.packages(c("ggplot2", "forecast"))
source("code_Time_Series.R")
```

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

Data and Getting Started

The data set for this example contains annual sales and advertising dollars (in 1960 dollars) for the Lydia Pinkham tonic from 1907 to 1960.These series are stored in two different objects called `tonic_sales` and `tonic_advert` , respectively.

Different advertising strategies were adopted from 1907-1914 (strategy A), 1915-1925 (B), 1926-1940 (C), and 1941-1960 (D). The advertising strategy is included using three dummy variables, `tonic_strat_B` , `tonic_strat_C` , and `tonic_strat_D` , with strategy A being the base case.

Also included are the first lags of sales and advertising in the series `tonic_lag_sales` and `tonic_lag_advert` . In all of seven of these series, the observation for 1907 has been removed to make all the series the same length due to the `NA` 's introduced by these lags.

```
load("data_Time_Series.Rdata")
```

```
# tonic_sales
# tonic_advert
# tonic_strat_B
# tonic_strat_C
# tonic_strat_D
# tonic_lag_sales
# tonic_lag_advert
```

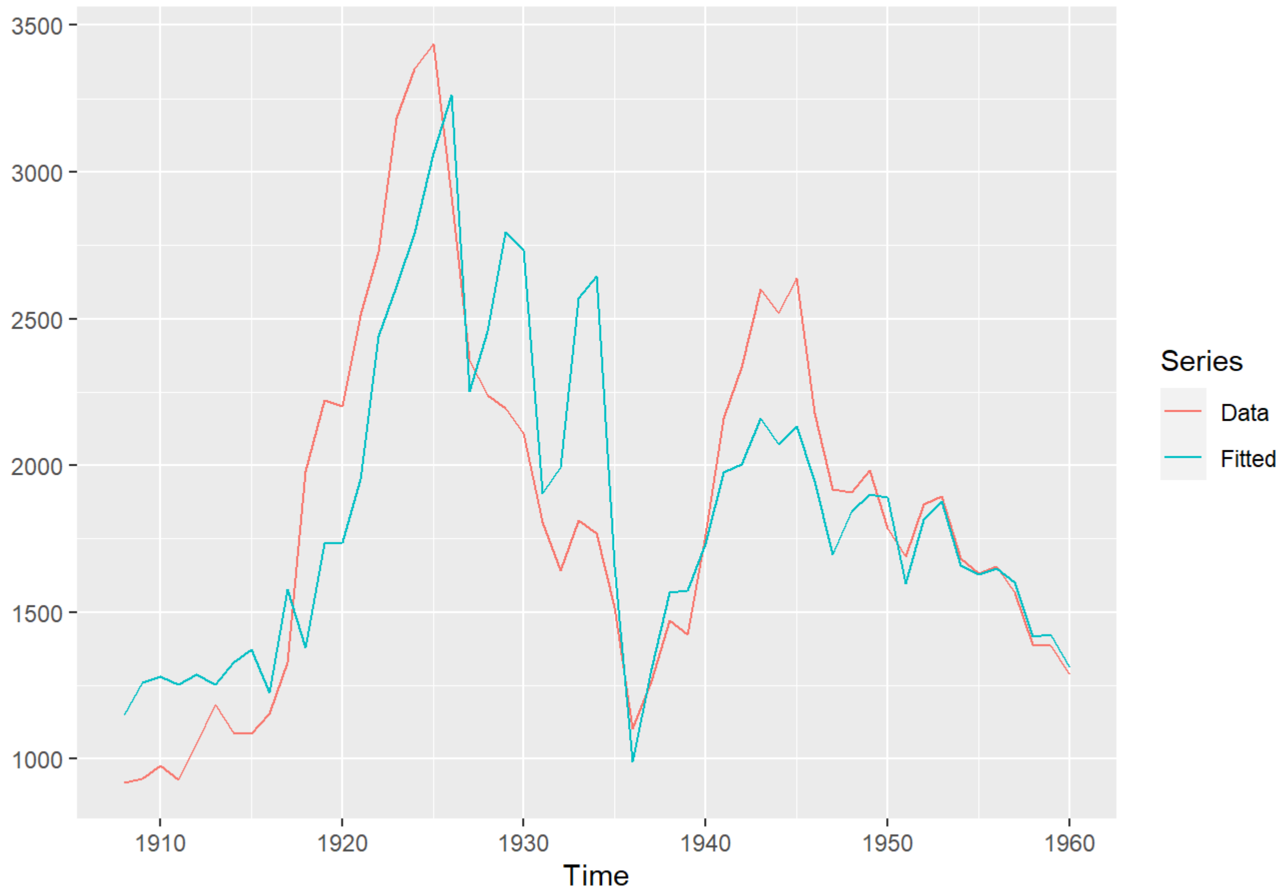
Used the `autoplot(...)` to print each of the series one by one. Reviewed to see if advertising and sales seem correlated. Whether advertising follows sales, or do sales follow advertising.

```
# autoplot(tonic_sales)
# autoplot(tonic_advert)
# autoplot(tonic_strat_B)
# autoplot(tonic_strat_C)
# autoplot(tonic_strat_D)
# autoplot(tonic_lag_sales)
# autoplot(tonic_lag_advert)
```

Just Sales and Advertising

Let’s start by first regressing `tonic_sales` on `tonic_advertising` to analyze the fit.

```
fit <- tslm(tonic_sales ~ tonic_advert)
aa_plot_fitted(fit)
```



The fit doesn’t look so great. Let’s also check the diagnostics.

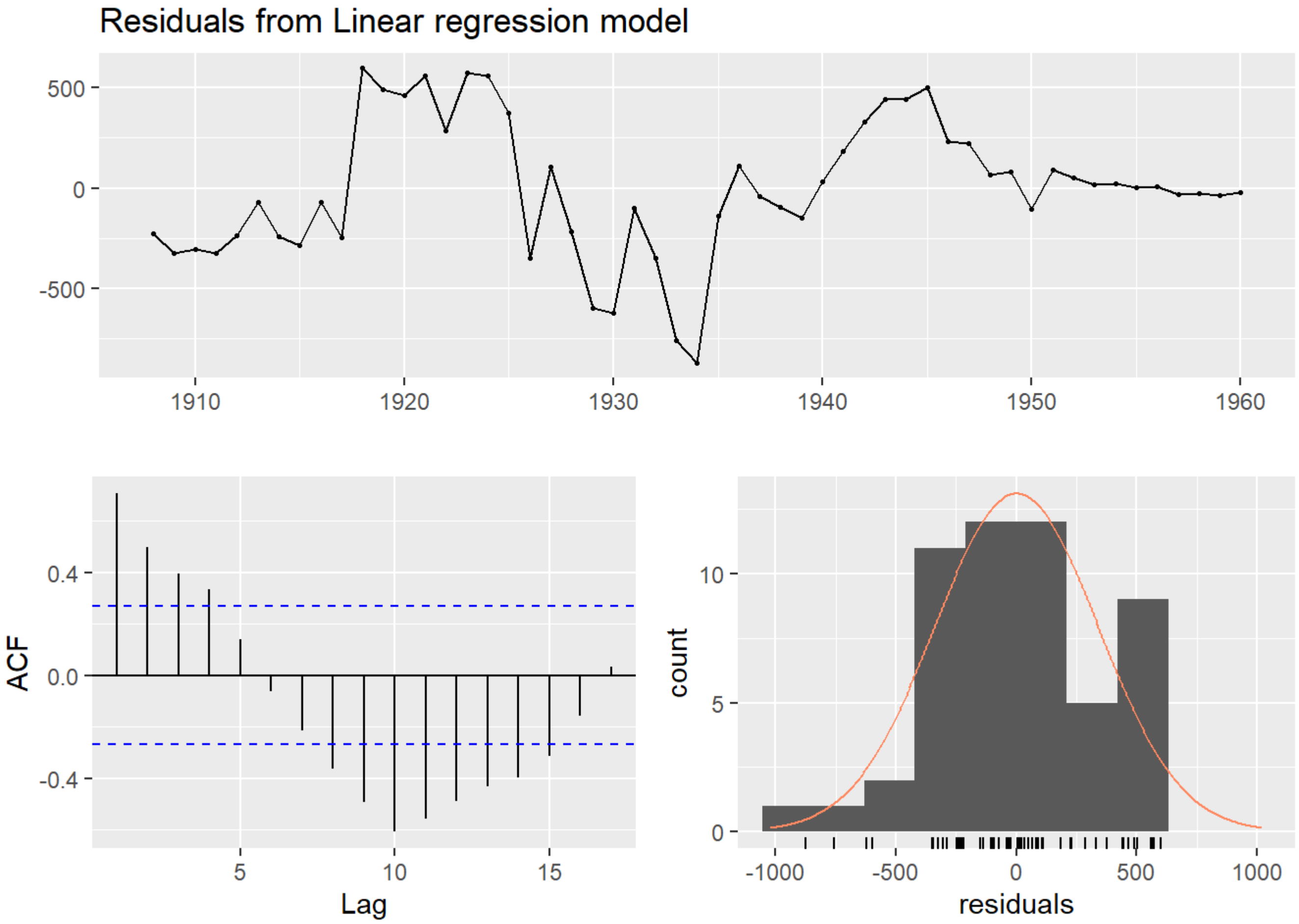
```
accuracy(fit)
```

```
##                        ME     RMSE      MAE       MPE     MAPE    MASE
## Training set 8.573396e-15 336.7783 258.7706 -3.439857 14.21672 1.45032
##                   ACF1
## Training set 0.7056093
```

```
summary(fit)
```

```
##
## Call:
## tslm(formula = tonic_sales ~ tonic_advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -874.01 -229.13  -21.44  223.68  600.05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  510.2978   129.1037   3.953 0.000239 ***
## tonic_advert   1.4187     0.1278  11.104 3.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 343.3 on 51 degrees of freedom
## Multiple R-squared:  0.7074, Adjusted R-squared:  0.7017
## F-statistic: 123.3 on 1 and 51 DF,  p-value: 3.218e-15
```

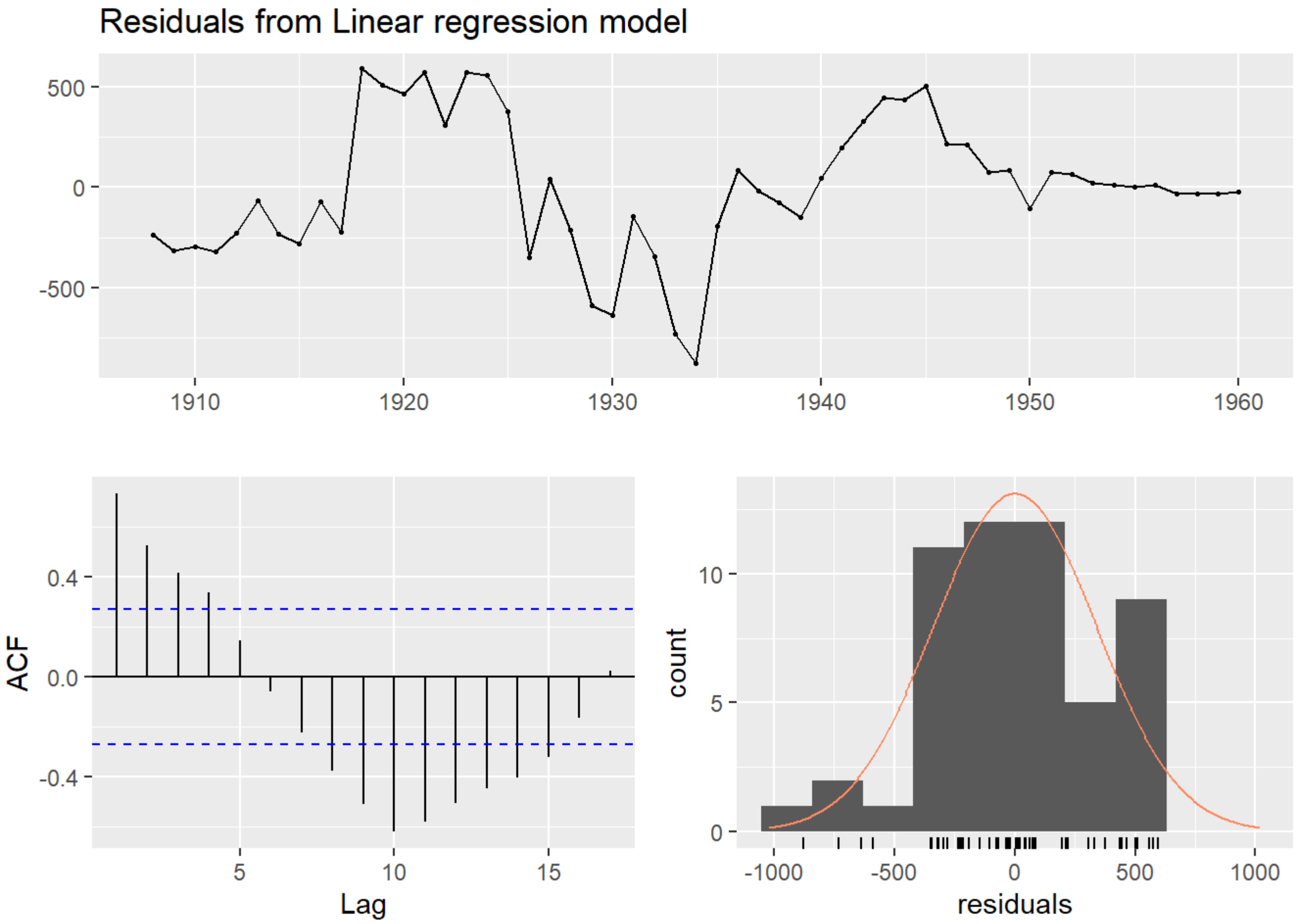
```
checkresiduals(fit)
```



```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals from Linear regression model
## LM test = 34.362, df = 10, p-value = 0.0001603
```

From above, it’s clear we are having trouble with autocorrelation. Maybe adding the lag 1 of advertising will help.

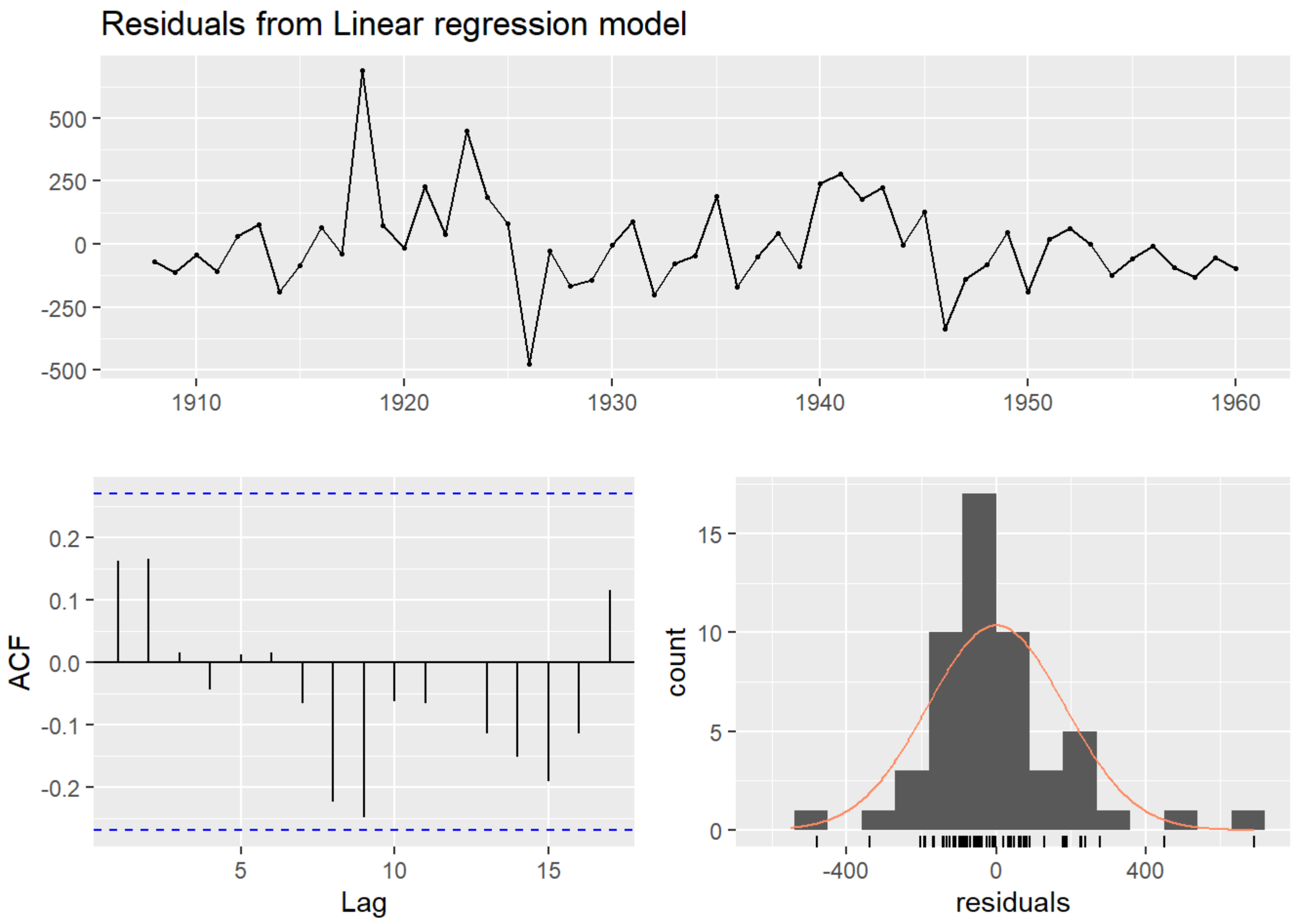
```
fit <- tslm(tonic_sales ~ tonic_advert + tonic_lag_advert)
checkresiduals(fit)
```



```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals from Linear regression model
## LM test = 40.979, df = 10, p-value = 1.138e-05
```

That didn’t seem to help the autocorrelation at all. Will now add lag 1 of sales to see if that helps improve the model.

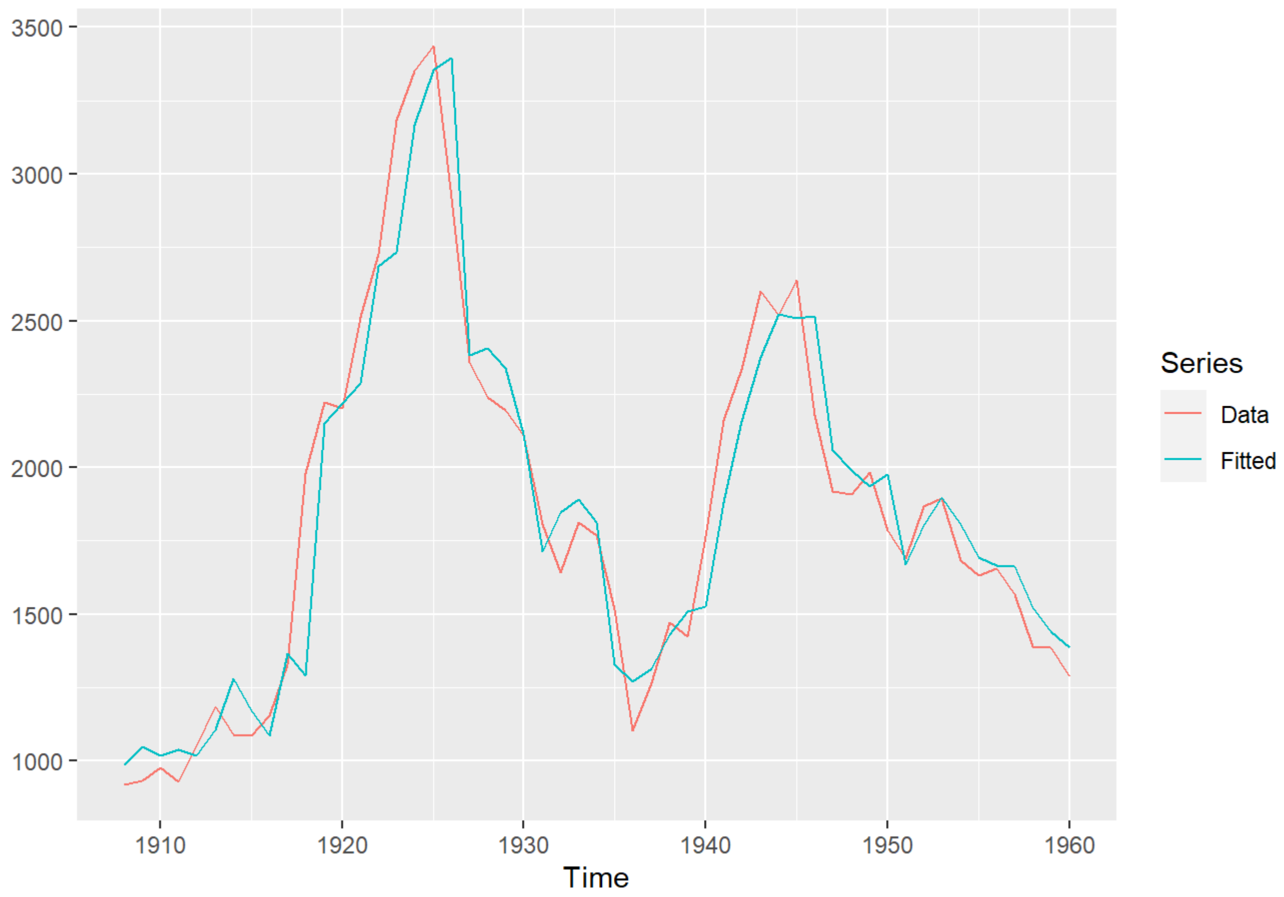
```
fit <- tslm(tonic_sales ~ tonic_advert + tonic_lag_advert + tonic_lag_sales)
checkresiduals(fit)
```



```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals from Linear regression model
## LM test = 9.1205, df = 10, p-value = 0.5207
```

Clearly that is much better! Let’s verify the fit visually:

```
aa_plot_fitted(fit)
```



Looks pretty good. Finally, let’s look at accuracy and verify variable significance:

```
accuracy(fit)
```

```
##                        ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 1.286654e-14 181.7368 129.2448 -1.048323 7.118487 0.7243728
##                  ACF1
## Training set 0.162151
```

```
summary(fit)
```

```
##
## Call:
## tslm(formula = tonic_sales ~ tonic_advert + tonic_lag_advert +
##     tonic_lag_sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -477.94  -97.66  -25.39   73.64  690.21
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      154.06533   80.49015   1.914 0.061458 .
## tonic_advert       0.58944    0.14232   4.142 0.000136 ***
## tonic_lag_advert  -0.66006    0.14156  -4.663 2.43e-05 ***
## tonic_lag_sales    0.95546    0.08764  10.902 1.06e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 189 on 49 degrees of freedom
## Multiple R-squared:  0.9148, Adjusted R-squared:  0.9096
## F-statistic: 175.4 on 3 and 49 DF,  p-value: < 2.2e-16
```

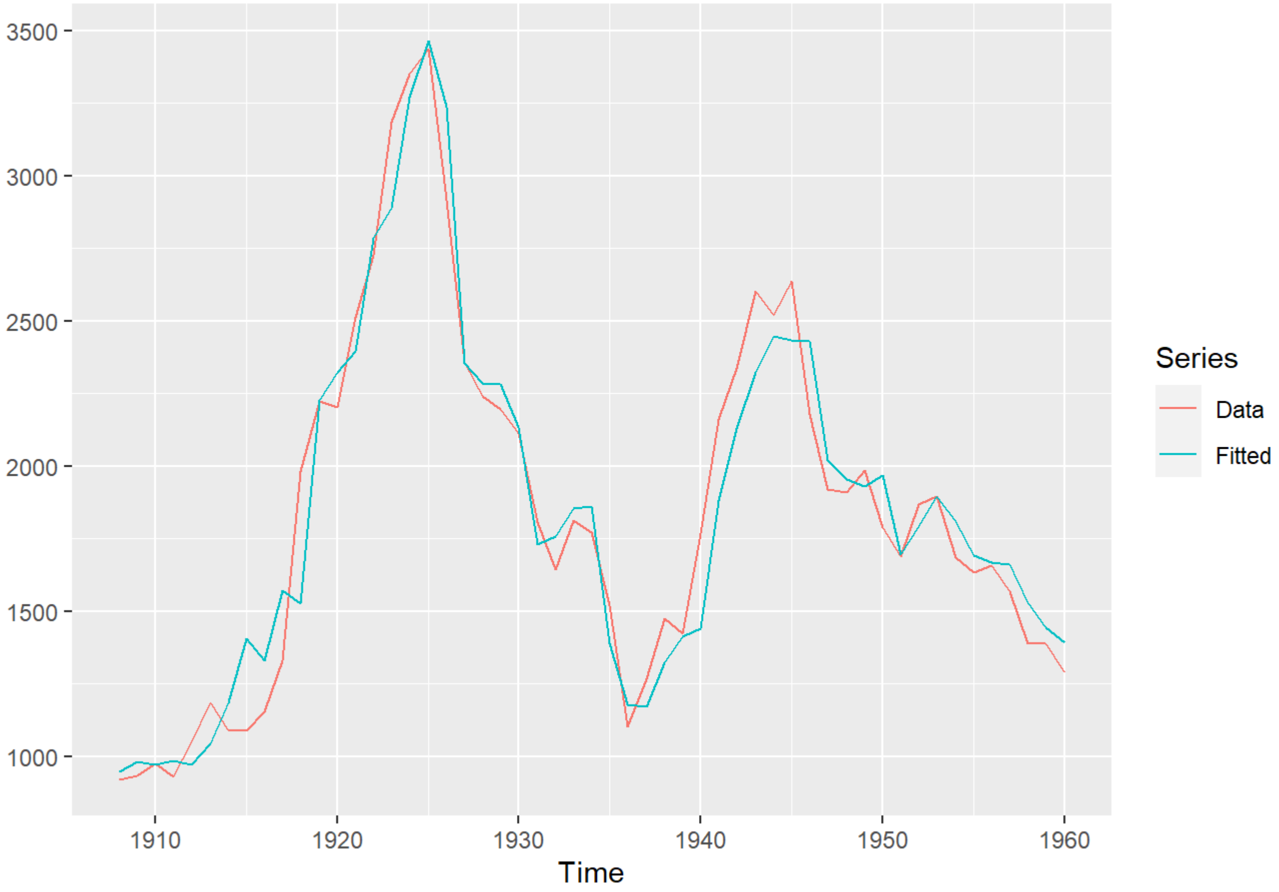
The MAPE is about 7%, and all variables are significant. This seems like a good model.

Adding in the Advertising Strategy

Now adding in the advertising strategy (without trying to remove insignificant variables):

```
fit <- tslm(tonic_sales ~ tonic_advert + tonic_lag_advert + tonic_lag_sales +
    tonic_strat_B + tonic_strat_C + tonic_strat_D)
```

```
aa_plot_fitted(fit)
```



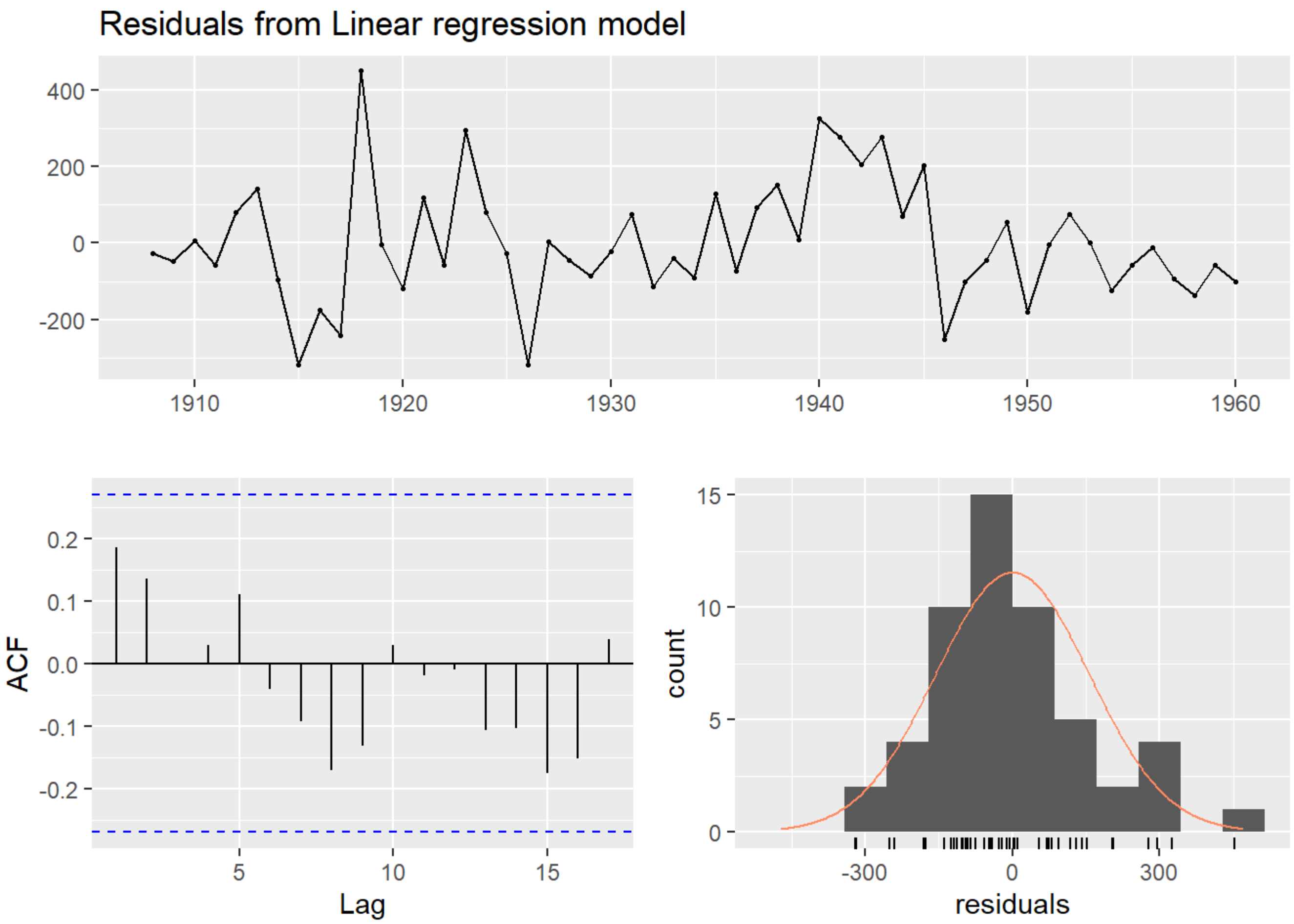
```
accuracy(fit)
```

```
##                        ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 1.286445e-14 155.6047 118.0198 -0.818896 6.804955 0.6614604
##                   ACF1
## Training set 0.1851649
```

```
summary(fit)
```

```
##
## Call:
## tslm(formula = tonic_sales ~ tonic_advert + tonic_lag_advert +
##     tonic_lag_sales + tonic_strat_B + tonic_strat_C + tonic_strat_D)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -319.81  -93.22  -26.86   79.42  452.28
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      128.87047   76.59058   1.683  0.09923 .
## tonic_advert       0.60279    0.13183   4.572 3.63e-05 ***
## tonic_lag_advert  -0.37913    0.15127  -2.506  0.01580 *
## tonic_lag_sales    0.76540    0.09968   7.678 8.85e-10 ***
## tonic_strat_B    296.46048   95.51591   3.104  0.00326 **
## tonic_strat_C    -11.86017   90.80719  -0.131  0.89665
## tonic_strat_D    104.99736   85.09806   1.234  0.22353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167 on 46 degrees of freedom
## Multiple R-squared:  0.9375, Adjusted R-squared:  0.9294
## F-statistic: 115.1 on 6 and 46 DF,  p-value: < 2.2e-16
```

```
checkresiduals(fit)
```



```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals from Linear regression model
## LM test = 6.9855, df = 10, p-value = 0.7268
```

All in all, this model looks quite good, and its error measures are better than the previous one without the strategies.

Final Thought: Forecasting

Since the model for forecasting `tonic_sales` is based on other time series, we have to forecast those time series first.

Let's pretend it's 1960 (the last year in the data), and we want to forecast `tonic_sales` for 1961.

Here's the sales amount from 1960:

```
window(tonic_sales, start = 1960, end = 1960)
```

```
## Time Series:
## Start = 1960
## End = 1960
## Frequency = 1
## [1] 1289
```

And here’s the advertising amount from 1960:

```
window(tonic_advert, start = 1960, end = 1960)
```

```
## Time Series:
## Start = 1960
## End = 1960
## Frequency = 1
## [1] 564
```

Let’s suppose that, in 1961, we will stick with strategy D and we plan to spend 500 on advertising. Then we can predict `tonic_sales` in 1961 as follows using the above formula:

```
128.87047 +          # intercept
    0.60279 * 500 +  # contribution from advertising
   -0.37913 * 564 +  # uses lagged advertising from 1960!
    0.76540 * 1289 + # uses lagged sales from 1960!
    104.99736        # contribution from strategy D
```

```
## [1] 1308.034
```

From the model, the above prediction is for tonic sales in 1961 following strategy D.