

Statistical Methods for Discrete Response, Time Series, and Panel Data: Live Session 7

Devesh Tiwari

June 27, 2017

Main topics covered in Week 6

- Classical Linear Regression Model (CLM) for time series data (You will have to review CLM by yourself)
- Linear time-trend regression
- Goodness of Fit Measures (for Time Series Models)
- Time-series smoothing techniques
- Exploratory time-series data analysis
- Autocorrelation function of different time series

Required Readings

- CM2009: Ch. 5.1 – 5.3
- SS2016: Ch.2

Agenda

1. Announcement: No class next week. Attend the Wed class or Thursday makeup session.
2. Review questions (20 mins)
3. Inspecting automotive sales in the US (30 mins)
4. Linear time trend regression (30 mins)
5. Wrap up

Breakout Session 1: Review Questions (10 mins in breakout rooms + 10 minutes discussion)

1. What does it mean to decompose a time-series and what would we find when we decompose one?
2. What is Gaussian White Noise? Describe its properties and importance in time-series analysis
3. Time-series analysis requires us to create many models. What are the hallmarks of a “good” model?

Breakout Session 2: Inspecting automotive sales in the US (10 mins in breakout + 20 mins group discussion)

An article by Daniel Gross of slate.com (http://www.slate.com/articles/business/moneybox/2017/05/the_auto_industry_is_in_trouble_and_it_could_cause_bigger_problems.html) points out that auto-sales have “hit a wall” and that monthly sales in 2017 are less than what they were a year prior. Let’s dig a little deeper into this by analyzing data from the St. Louis Federal Reserve on total auto sales in the United States (<https://fred.stlouisfed.org/series/TOTALNSA>).

We will not directly address whether monthly auto sales have been declining compared to their prior year values. Instead, let’s examine whether or not there is evidence that automotive sales have been decreasing in 2017.

This is a monthly time-series data that is **not** seasonally adjusted!

1. Use either a moving average filter (filter) or a kernel smoother (ksmooth) to examine the underlying trend in this data. What do you notice about this data? Is there a trend? Seasonality?
2. For a given filter, what parameters did you have control over? When manipulating that parameter, what do you find?
3. Based on your examination, do you think that automotove sales have been decreasing? How would you go about examining this with a model?

```
rm(list = ls())
library(psych)

path <- "~/Documents/Projects/MIDS/Summer 2017/live_sessions/week7"
setwd(path)
df <- read.csv("vehicle_sales_NSA.csv", stringsAsFactors = FALSE)
head(df)
```

```
##          DATE TOTALNSA
## 1 1976-01-01    885.2
## 2 1976-02-01    994.7
## 3 1976-03-01   1243.6
## 4 1976-04-01   1191.2
## 5 1976-05-01   1203.2
## 6 1976-06-01   1254.7
```

```
tail(df)
```

```
##          DATE TOTALNSA
## 492 2016-12-01   1717.9
## 493 2017-01-01   1164.3
## 494 2017-02-01   1352.1
## 495 2017-03-01   1582.7
## 496 2017-04-01   1449.9
## 497 2017-05-01   1541.7
```

```
describe(df)
```

```
## Warning: NAs introduced by coercion
```

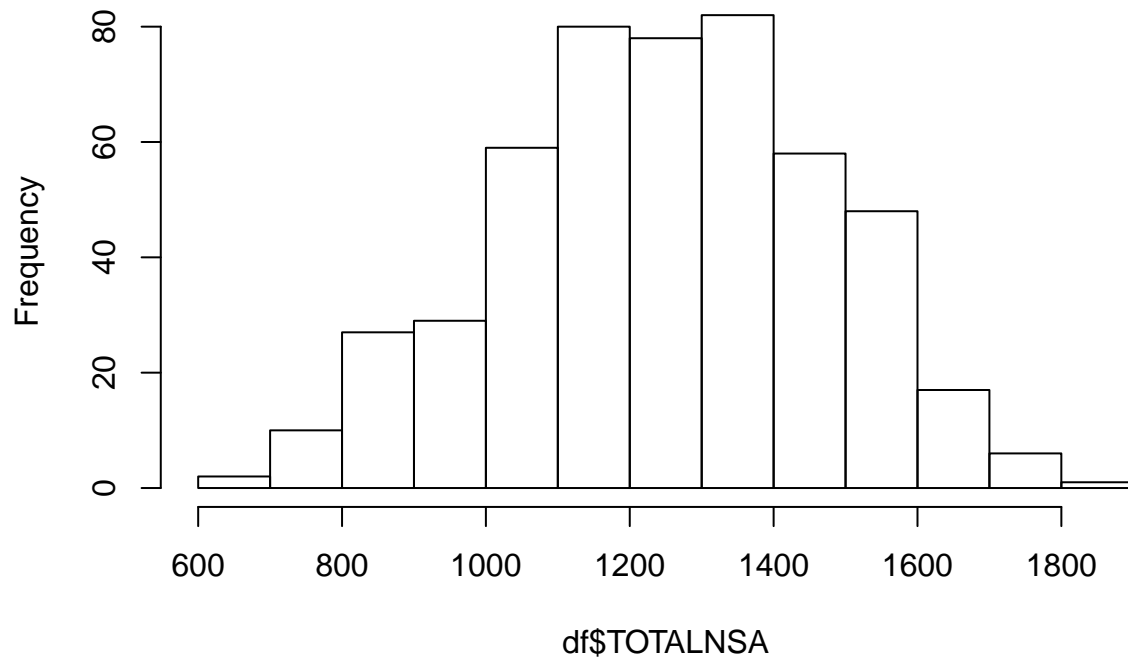
```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning
## Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning
## -Inf
```

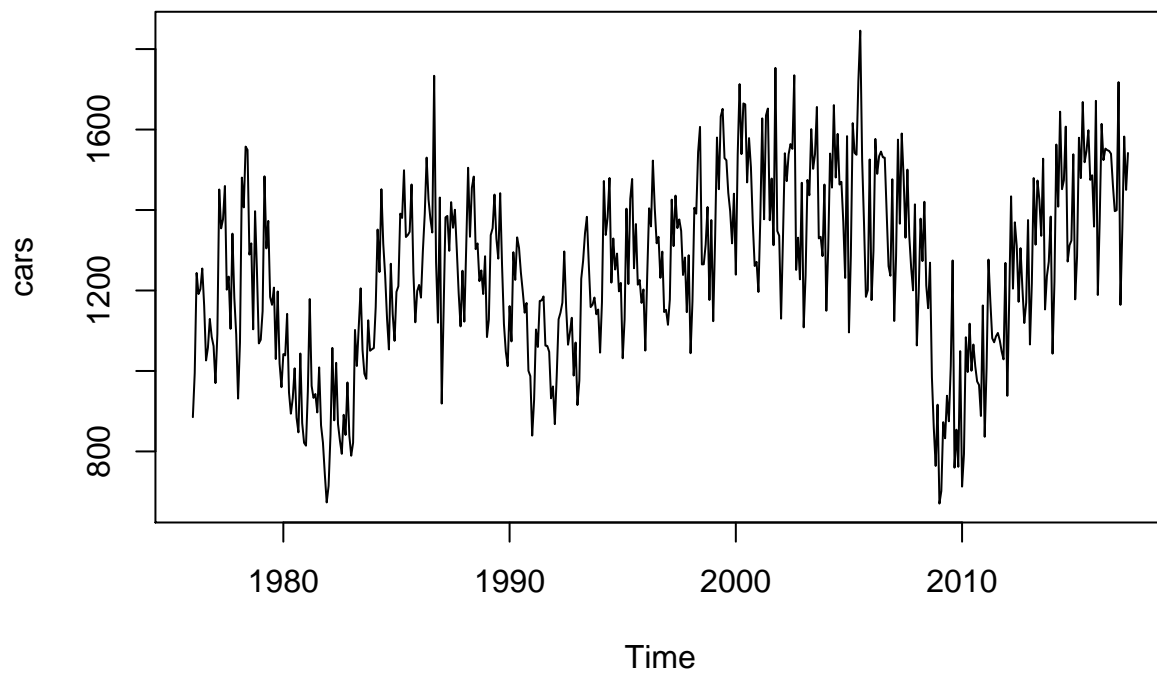
```
##          vars    n    mean      sd median trimmed   mad   min    max range
## DATE*         1 497     NaN     NA      NA      NaN    NA   Inf  -Inf  -Inf
## TOTALNSA      2 497 1249.11 223.76 1252.1 1254.02 231.73 670.4 1845.7 1175.3
##          skew kurtosis    se
## DATE*        NA        NA    NA
## TOTALNSA -0.14    -0.45 10.04
```

```
hist(df$TOTALNSA)
```

Histogram of df\$TOTALNSA

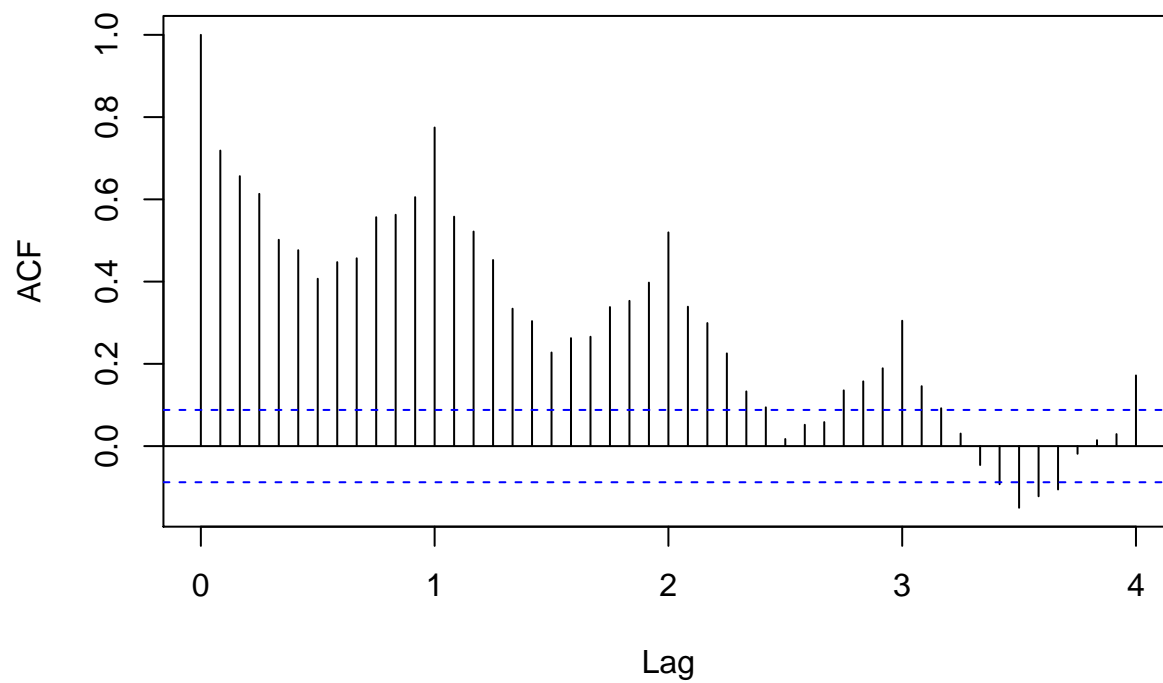


```
cars <- ts(df$TOTALNSA, frequency = 12, start = c(1976,1))
plot(cars)
```



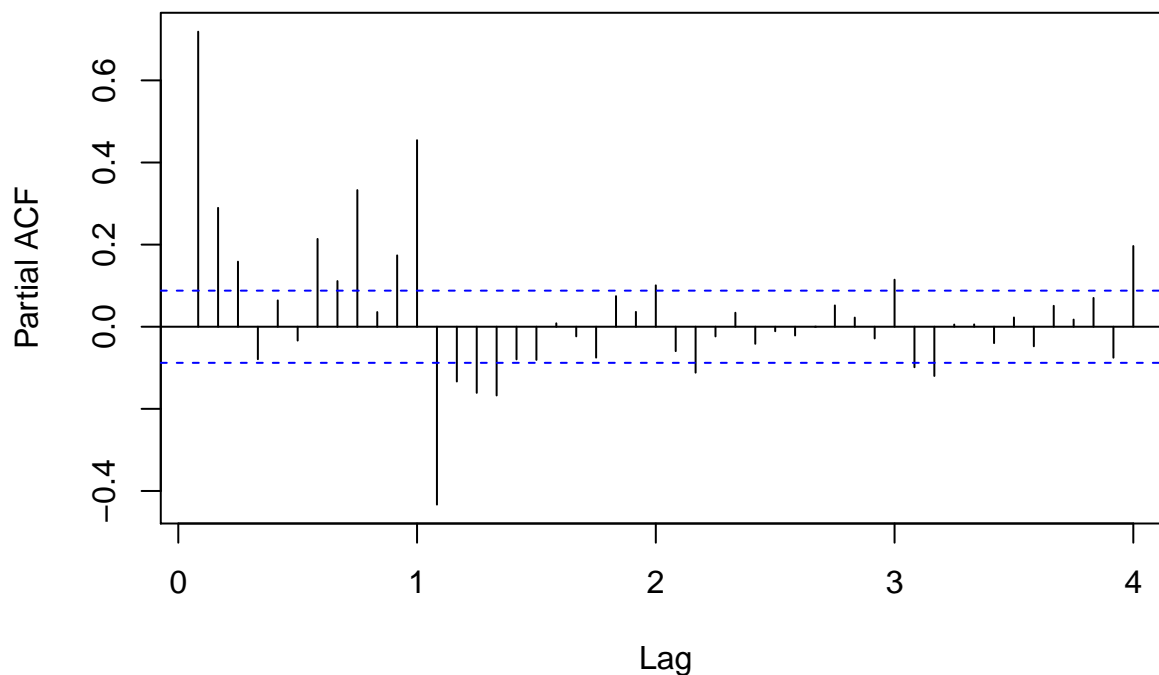
```
acf(cars, lag.max = 48)
```

Series cars



```
pacf(cars, lag.max = 48)
```

Series cars



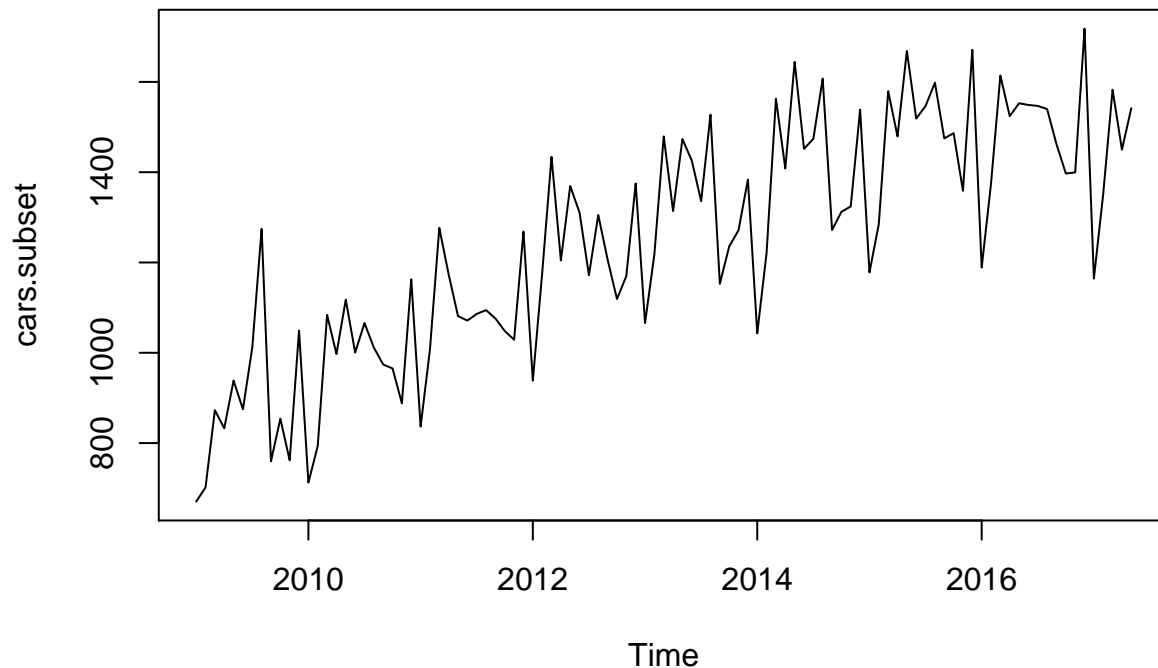
Insert code here for moving average or kernel filters

Breakout Session 3: Linear time regression (10 mins in breakout; 20 mins discussion)

Let's focus our attention on the time-period after 2009 in order to examine the trend in automotive sales. In the code below, I conduct an OLS regression on the number of car sales and time; for this analysis, I include time as a quadratic term. Examine the output below and answer the following questions:

1. Based on your interpretation of the model, what is your interpretation about car sales in the US during this time period? Do you share Daniel Gross's concern that we need to be worried about the US economy?
2. Examine the residuals. Do they resemble gaussian white noise? Based on your residuals analysis, does your interpretation of the results above change?
3. What other changes would you make to this model?

```
## Subset the time-series
cars.subset <- window(cars, start = c(2009,1))
plot(cars.subset)
```



```
sales <- as.numeric(cars.subset)
time <- as.numeric(time(cars.subset))

model1 <- lm(sales ~ time + I(time^2))
summary(model1)
```

```
##
## Call:
## lm(formula = sales ~ time + I(time^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -336.09  -97.99   10.47   94.84  402.62
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.609e+07  1.065e+07  -3.389  0.00101 **
## time         3.577e+04  1.058e+04   3.381  0.00104 **
## I(time^2)    -8.863e+00  2.628e+00  -3.373  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139.4 on 98 degrees of freedom
## Multiple R-squared:  0.7129, Adjusted R-squared:  0.707
## F-statistic: 121.7 on 2 and 98 DF,  p-value: < 2.2e-16
```

```
#INSERT CODE HERE FOR FURTHER ANALYSIS, RESIDUALS ANALYSIS, ETC.
```

Overview of modelling time-series data

Over the next few weeks, we will learn how to analyze time-series data using methods that treat the current observation as some function of its prior observations. Next week, we will talk about two types of processes, the AR(p) and MA(q). As mentioned before, we will learn how to handle more complicated data, which in turns makes the modelling process a little more complicated. Having said that, the broad outlines of the modeling process are the same:

1. EDA During the EDA phase, we want to get a sense of how the time-series data behaves. In particular, we want to examine in if we need to transform the data in anyway in order to make it suitable for analysis. All of our methods require us to model data that are *weakly stationary*. This is where we determine if the time-series fits that criteria.

We also examine the ACF and PACF plots in order to form our initial guesses as to how we should model the data.

2. Model building Once we have transformed the data, we have to model it. At this stage, we need to determine the following:
 - (a) Do we need to include an AR component? an MA compent?
 - (b) How many lags of each should we include?

It should be apparent that we will estimate *many* models in this stage. So our task here is selecting a handful models that we think are good candidates. At this stage, we typically use a combination of ICs and residuals analysis to make that determination.

3. Model evaluation: Out of sample fit Once we have a handful of models from step 2, we compare them head to head based on their predictive accuracy. We then select one of these models as *our* model.
4. Generate forecasts and answer the question Once we have a final model, we should generate some forecasts. Sometimes, you might want to generate forecasts using more than one model. While you are free to do so, at the end of a modeling exercise, you should select one model to be your chosen model.

Final note: There are functions in R that can select a model for you. *DO NOT* use them in this class for purposes other than checking or validating your own work. Even then, be sure to read the documentation so you understand how the function is selecting a model.