MATH 645- Project 1

Fall 2019

---

1 Abstract

1) Summary of problems

- Describe the attributes of New Private Housing Units Authorized by Building Permits for Colorado.
- Describe how the response variable for Colorado compare to the same variable at the national level.
- Describe the attributes of the data for each of the explanatory variables I considered.
- Describe the association between the response variable for Colorado and the explanatory variables I considered.

2) Summary of methods

- Method for describing the attributes of New Private Housing Units Authorized by Building Permits for Colorado:
  ACF , Linear Regression, Residual Analysis, Detrending and Differencing, Smoothing, Holt-Winters Method, Forecasting.
- Method for describing how the response variable for Colorado compare to the same variable at the national level:
  CCF, Regression with Autoregressive Errors, Data transformation
- Method for describing the attributes of the data for each of the explanatory variables I considered:
  Detrending, Differencing, Smoothing, Linear Regression
- Method for describing the association between the response variable for Colorado and the explanatory variables I considered:
  Linear Regression, ANOVA, Cross Validation, Stepwise Regression.

2 Introduction

1) Statement

- The objective of the project is to evaluate the state of new housing construction at the state level, which is Colorado. Due to the result we derive with several methods, we detect the comparison with the same variable at the national level. Then, we detect the features and attribution of the explanatory variables. Finally, we dig out the association between the response and the explanatory.

2) Clarification

- When I get the first-hand raw data, how can we fetch and manage the data to be well prepared for the following research?
- How can I summary the result I get from the method and how do they work and help me to research?
- When I have already detected the features of the data, is it possible to do the forecast? Is the forecast what we really expect?
- If the association is hard to evaluate, how can I transform the data to remove the effect on which some 'bad' things could have?
- How can I present a clear visualization of the data?

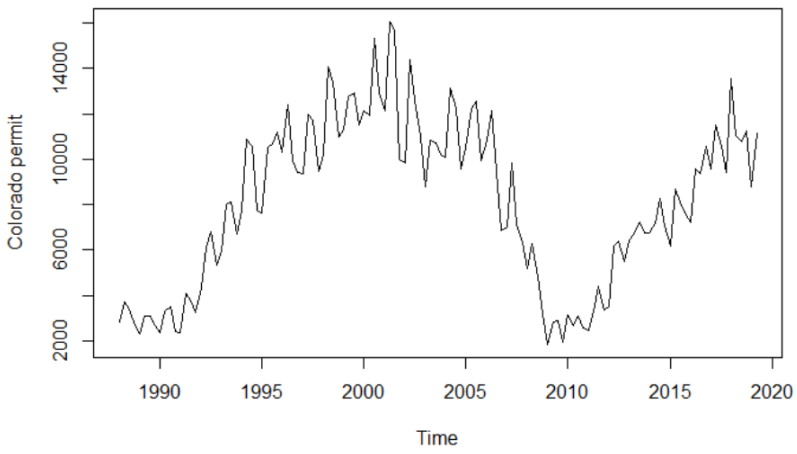
3 Method and Materials

1) Resource of response and explanatory variables

- **Response Variable:**
  COBP: New Private Housing Units Authorized by Building Permits in Colorado
  PER: New Private Housing Units Authorized by Building Permits in the USA
  Reference: FRED
- **Explanatory Variable:**
  LFP: Labor force participation in Colorado
  COUR: Unemployment Rate in Colorado
  DIR: Dividends, Interest and Rent in Colorado
  Reference: state level data on FRED

2) Summary of individual variable property

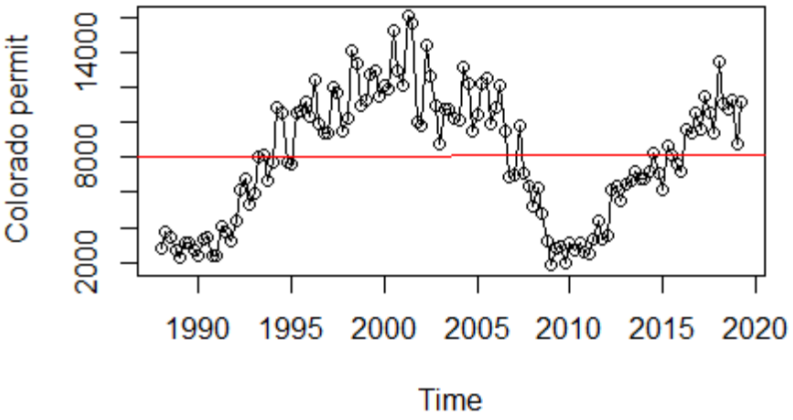
- **Q1: Describe COBP**
  It is the new private housing units authorized by building permits in Colorado which is collected between 1988 and 2019 after data fetching.



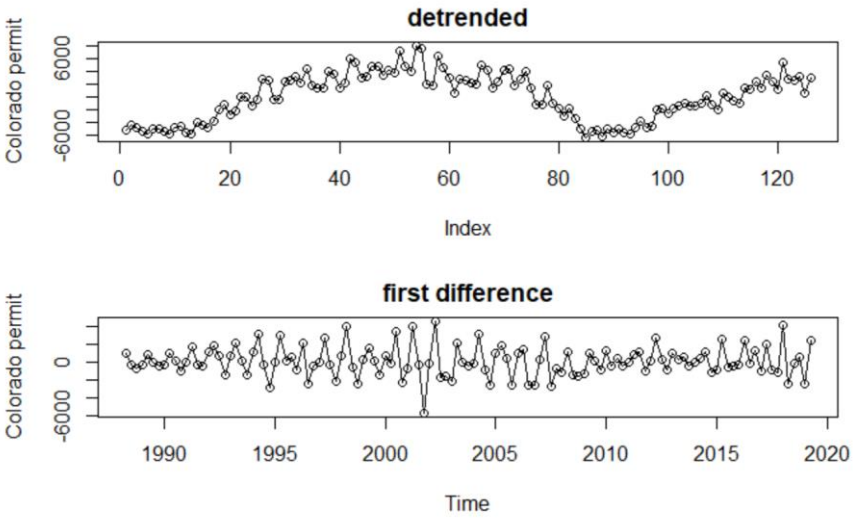
From the plot, we see that there are frequent fluctuations during these years. Also, there seems to be a peak at around 2002 and drop down to the bottom at around 2009. It really makes sense since the financial crisis broke out at that time. In other

words, it will drive people change the capital structure and investment demand, which I assume will drive the COBP down. But it kept going up till now after the recovery at around 11000.

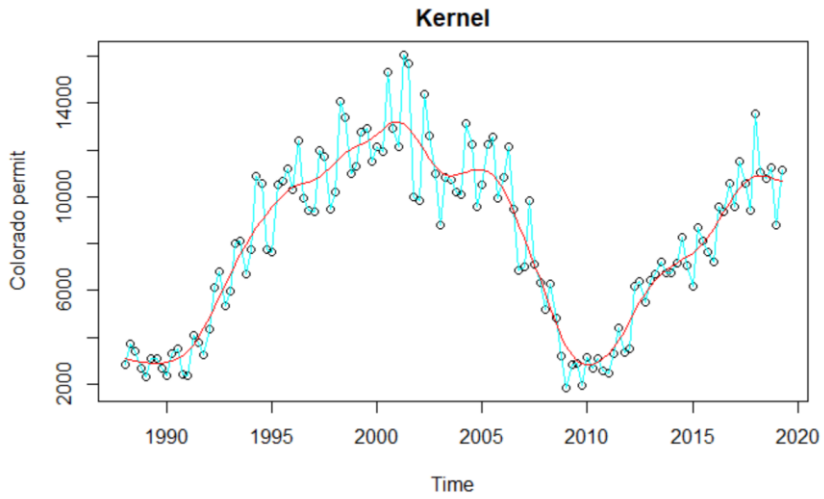
Then I use describe function in R to have a brief summary of the data and I find that the mean of COBP is 8094.7, the standard deviation is 3666.13 and the range is 14222. Also, I use ACF function in R to detect the autocorrelation of the data, which reflects relatively strong autocorrelation when lag is around 4. Then, I want to detect the attribute of COBP with time shifts by regression analysis

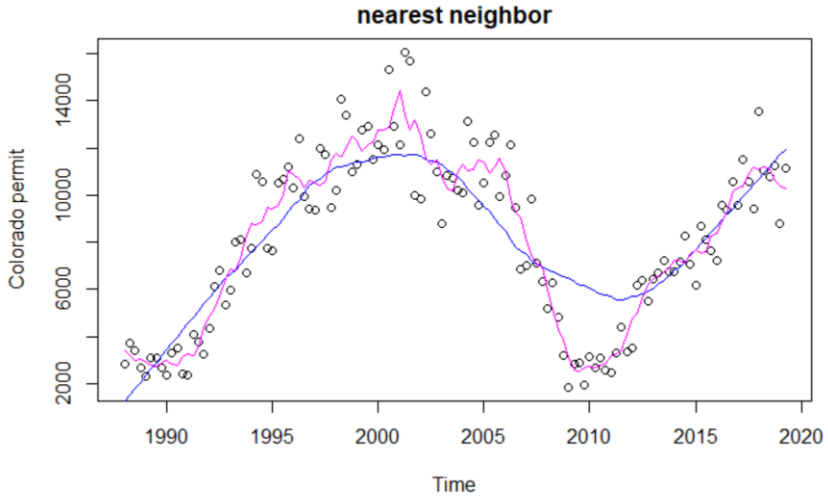


After centering the data, I find that COBP fluctuate at the mean which is 8000. And the slope of the fitted value is flat and it does not give me a clear association. Let me detrend the data with the residual and first difference. And I think the latter performs better to detrend.

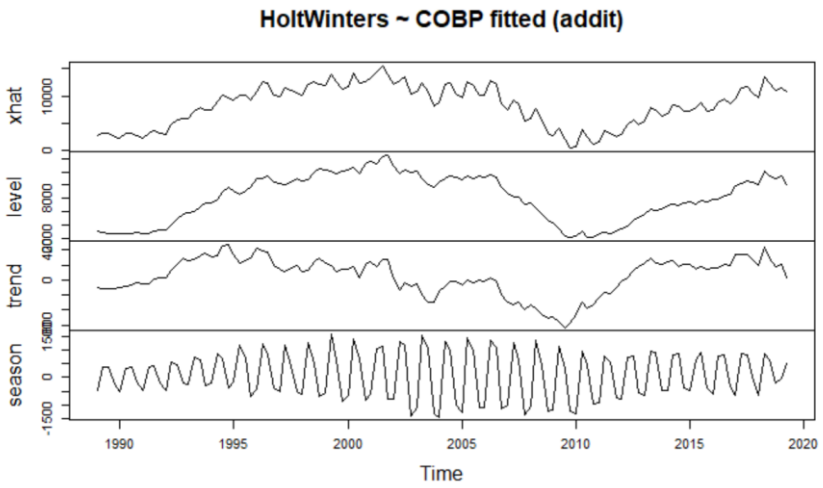


Then, I use 2 smoothers to smooth the data which are Kernel Smoother and Nearest Neighbor Regression Smoother.

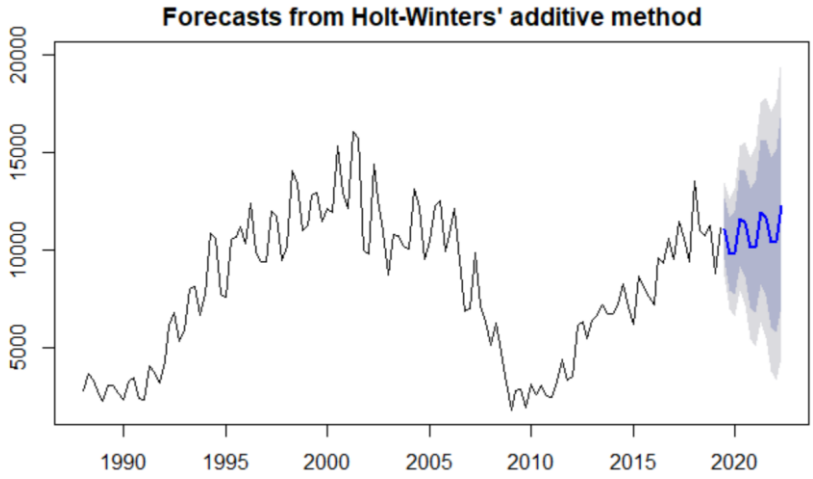




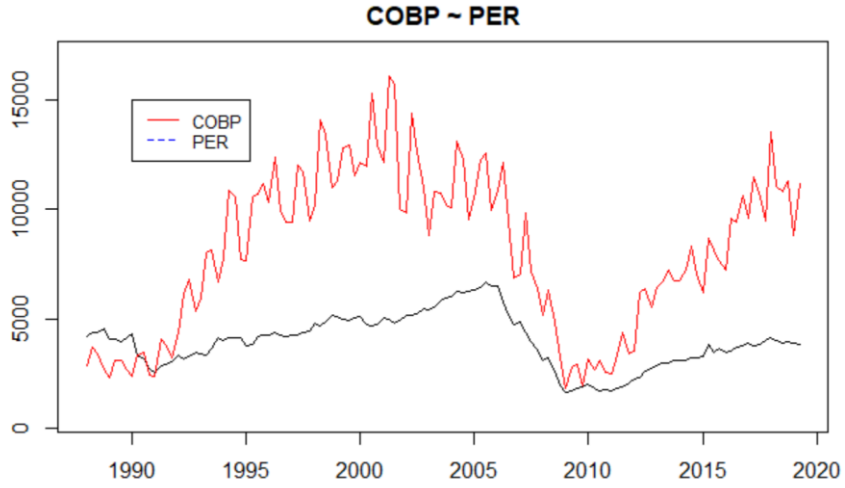
Also, I use Holt-Winters Method (multi) which is exponential smoothing strategy to estimate the level, slope and seasonal component at the current time point. The estimated values of alpha, beta and gamma are 0.49, 0.13, and 0.30, respectively. The value of alpha (0.49) is relatively low, indicating that the estimate of the level at the current time point is based upon both recent observations and some observations in the more distant past. The value of beta is 0.13, indicating that the estimate of the slope b of the trend component is not updated over the time series, and instead is set equal to its initial value. This makes good intuitive sense, as the level changes quite a bit over the time series, but the slope b of the trend component remains roughly the same. Also, the value of gamma (0.30) is low, indicating that the estimate of the seasonal component at the current time point is nearly not based upon very recent observations.



Then, I do the short-term forecast in 12 months.

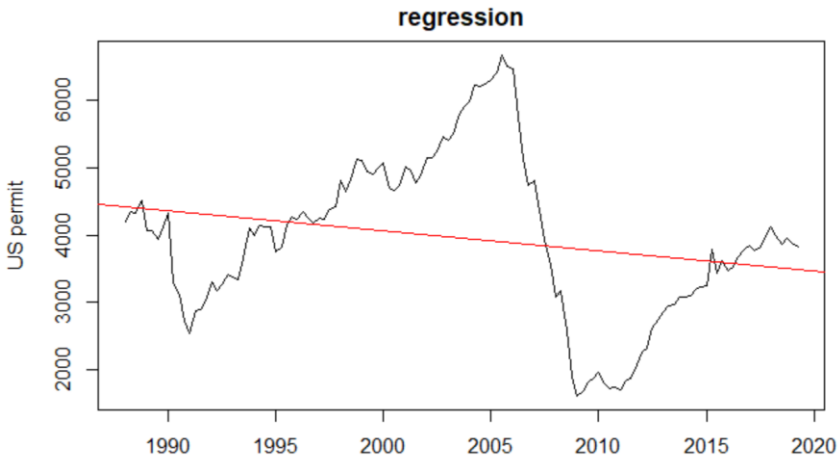


- **Q2: Compare COBP with PER**



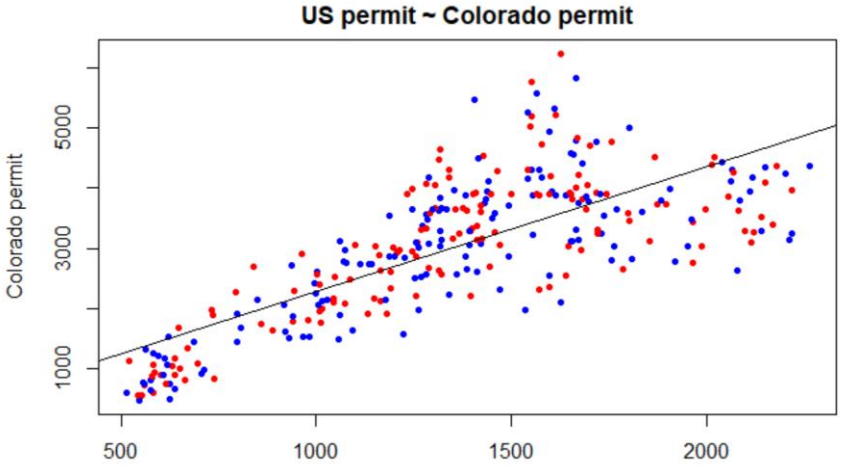
From the graph, we can see clearly that except for 1990s, COBP is higher than PER generally. And the PER reached the peak at around 2005 instead of 2002. And PER also drop down to the bottom at around 2009 for the financial crisis.

Then, I did the regression analysis on PER.



The graph gives me a downward trend which I think the response variable in the national level (PER) is affected more severely by financial crisis than the same one (COBP) was. But I think it will be recovered as well. So, why not forecast it by regression with autoregressive errors?

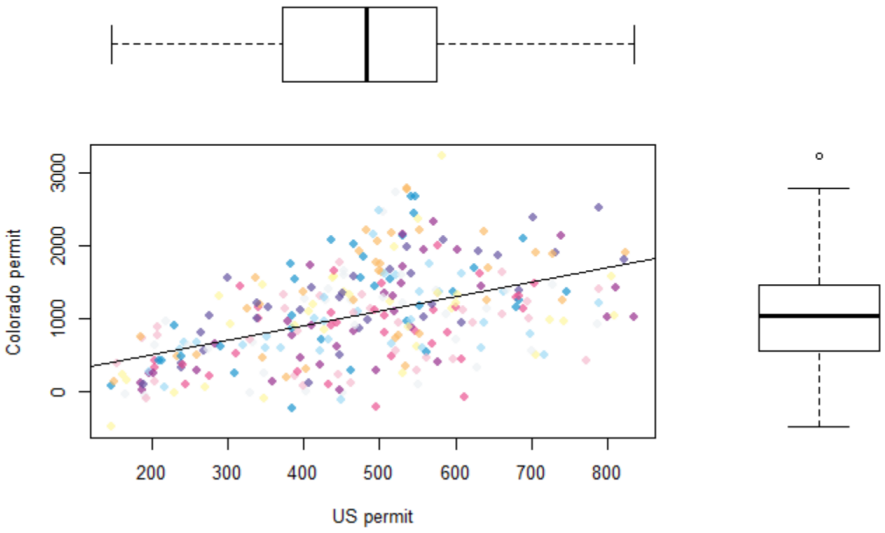
So, I extract the PER and COBP for 300 rows and combine them together and use lm function in R to detect the linear relation.



Since we know that (from 645 note3)

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_{t-1} e_t}{\sum_{t=2}^{n} e_{t-1}^2}$$

After computing rho and transforming the data (R code is attached). We find the transformed β to be 315.8447 and 1.992542. Then, I plot the transformed data.



Since I can obtain an approximate (1-α)% prediction limit for F:
(from 645 note3)

$$F_{n+1} \pm t_{\frac{\alpha}{2},n-3} SE_{prediction}$$

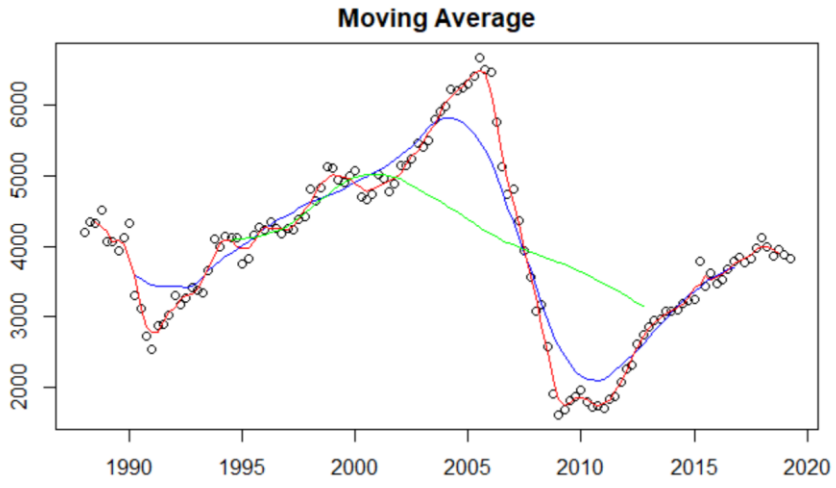
where

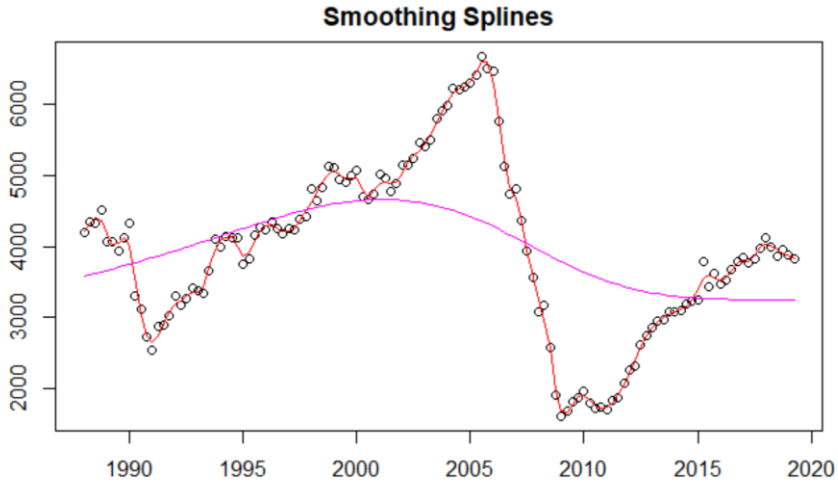
$$SE_{prediction} = \sqrt{MSE(1 + \frac{1}{n} + \frac{(X'_{n+1} - \bar{X}')^2}{\sum_{i=1}^{n} (X'_i - \bar{X}')^2})}$$

I compute the prediction interval to be (2278.386, 4605.645) when α is 5%

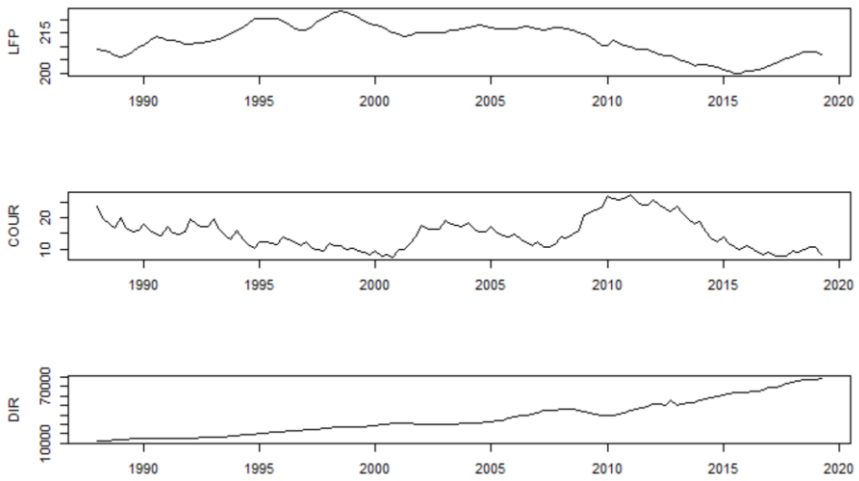
.

Then, I also use 2 different smoothers to smooth PER which are Moving Average Smoother and Smoothing Splines (from my R code).



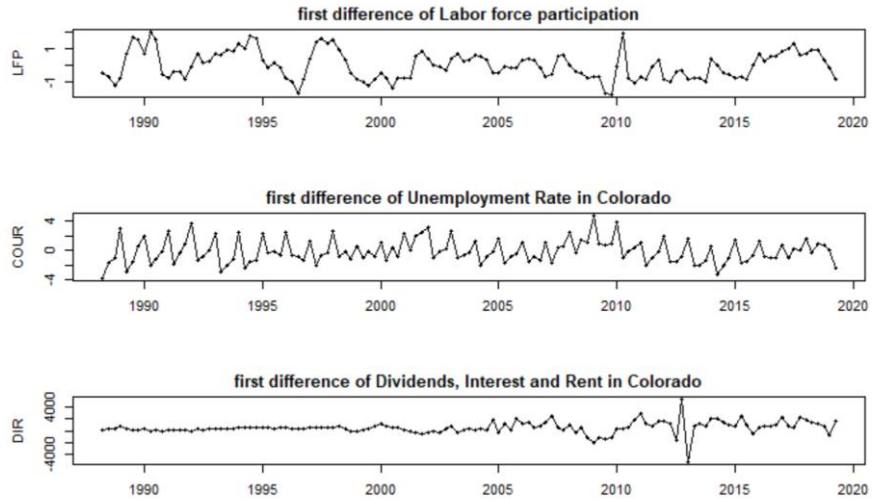


- **Q3: Describe LFP, COUR and DIR**

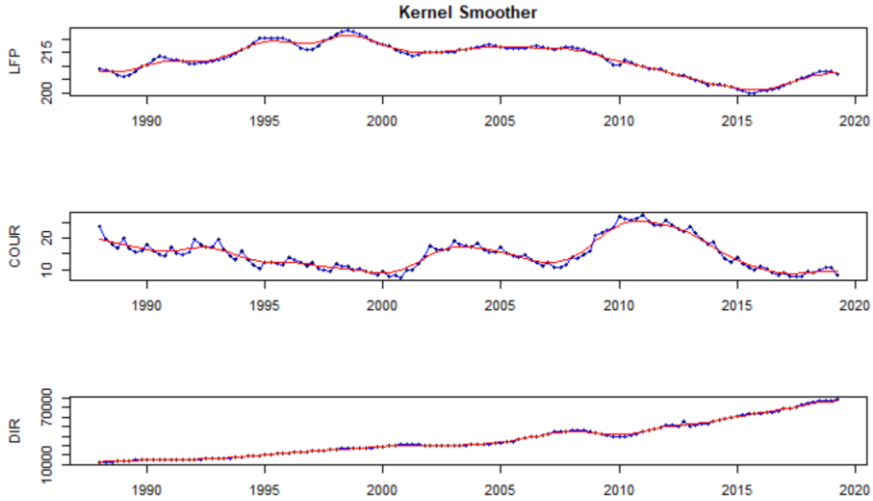


From the plot, I find that LFP and COUR go up and down, whereas DIR seems to be growing during these years. And when we cut to the time point at 2009, it seems that the Labor Force Participation (LFP) was not badly affected by financial crisis. However, the Unemployment rate (COUP) almost reached the peak and Dividends, Interest and Rent dropped down a little bit.

As usual, I detrend the data by first differencing.



And I use Kernel Smoother to smooth each of these 3 variables.



- **Q4: Describe the association between COBP and LFP, COUR, DIR**

  Firstly, I use lm function in R to fit the multi-regression model and plot the result.

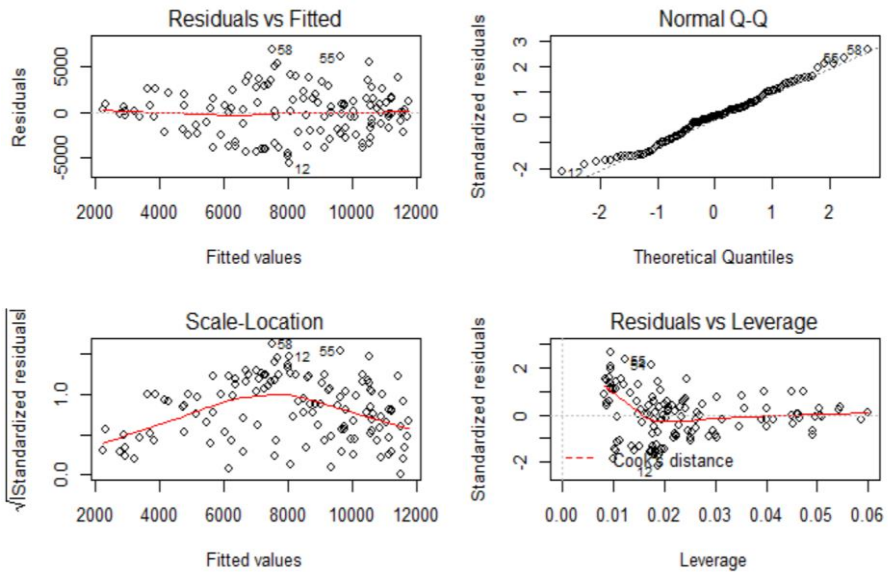
  Here are the result by calling lm function in R after centering the data.

```
Call:
lm(formula = y ~ x1 + x2 + x3 - 1, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-5606.2 -1963.7    36.2  1529.6  6883.0

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
x1   68.49141    4.15938  16.467   <2e-16 ***
x2 -475.57152   46.19196 -10.296   <2e-16 ***
x3    0.01818    0.01250   1.454    0.149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2622 on 123 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9128
F-statistic: 440.9 on 3 and 123 DF,  p-value: < 2.2e-16
```



4 Result and Conclusion

- The financial crisis significantly affected new private housing units authorized by building permits for Colorado and other explanatory variables. Nearly every variable has the feature that the value dropped down at around 2009 to the bottom.
- After fetching and managing the data, I detrend the data by first differencing, smoot the data by different smoothers and forecast PER by regression with autoregressive errors and forecast COBP with Holt-Winter Method in a short term. All of them fits not badly and I interpret what I do in the R code explicitly.
- In terms of multi-regression model, I find that 91.49% of variation in the PER can be explained by LFP, COUR and DIR. And PER is positively related to LFP with the slope to be 68.49, negatively related to COUR with the slope to be -475.57 and almost not related to DIR since the slope is close to 0. And I also use and interpret some other methods such as Cross validation, Stepwise regression in R code.