Code

Unemployment in the US: How to Predict and build a model of it

by Yaroslav Halonko, Andrian Hevalo

\(\textbf{Abstract}\)

Unemployment mostly falls during periods of economic stability and rises during recessions, creating significant pressure on public finances as tax revenue falls and social safety net costs increase. In this project We are trying to build a model of predicting unemployment rate for further years. Here we will use balanced panel macroeconomic data of the US. Although, our data has only indicators of the US rates, it is divided by quarters. This led us to the idea of checking the differences between them and providing interesting hypothesis testings.

\(\textbf{Short Contents}\)
- Visualizing Data
- Bulding bunch of models
- Checking for correctness of the models & diagnostics (and choosing the best model)
- Hypothesis testings
- Conclusions

Packages for installation & Usage
Here \(\textbf{car}\) package is for scatterplots, \(\textbf{gplots, lmtest, tseries}\) packages are for diagnostics of builded models, and \(\textbf{plm, foreign}\) packages are for bulding panel data models.

Hide

```
#install.packages('plm')
#install.packages('foreign')
#install.packages("gplots")
#install.packages("car")
library("car")
library("gplots")
```

```
package 惕柂gplots惕柞 was built under R version 3.5.3
Attaching package: 惕柂gplots惕柞

The following object is masked from 惕柂package:PerformanceAnalytics惕柞:

    textplot

The following object is masked from 惕柂package:stats惕柞:

    lowess
```

Hide

```
library('foreign')
library('plm')
library("lmtest")
library("tseries")
```

Reading CSV-file. Here we have the US macroeconomic data for the time period of 1950-2000.
Silght description of variables:
Year = Date
Qtr = Quarter
Realgdp = Real GDP in the US($bil)
Realcons = Real consumption expenditures
Realinvs = Real investment by private sector
Realgovt = Real government expenditures
Realdpi = Real disposable personal income
CPI_U = Consumer price index
M1 = Nominal money stock
Tbilrate = Quarterly average of month end 90 day t bill rate
Unemp = \(\textbf{Unemployment rate}\) (which we will try to model, our depended variable)
Pop = Population, mil. interpolate of year end figures using constant growth rate per quarter
Infl = Rate of inflation (first observation is missing)

Realint = Ex post real interest rate = Tbilrate - Infl. (First observation missing)

Hide

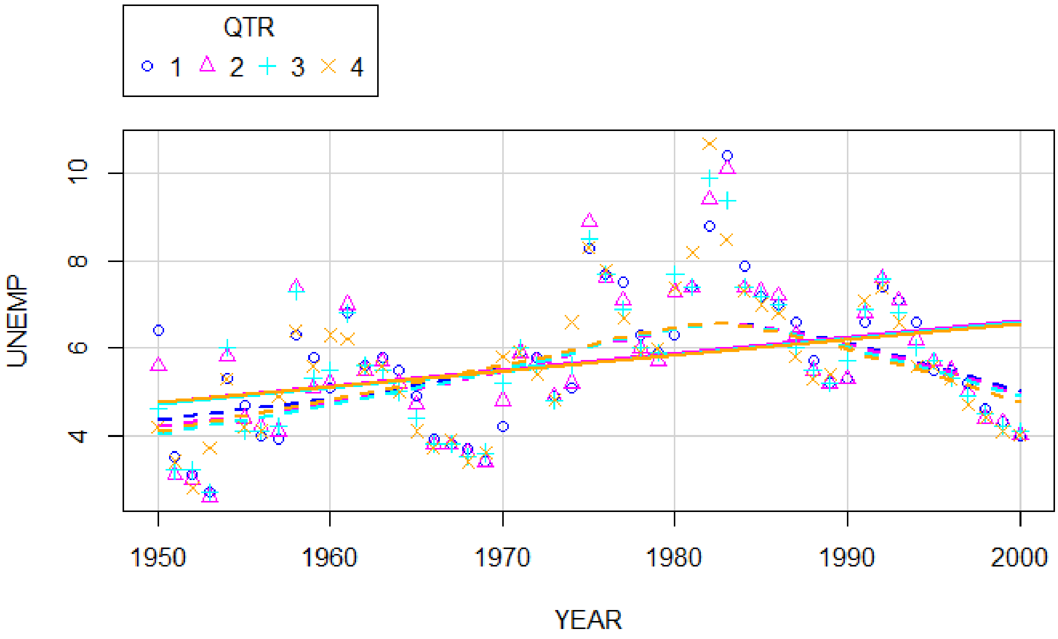
```
UN_data<- read.csv('D:/Rlabs/econ_project/Econometrics_project/TableF5-2.csv')
head(UN_data)
```

Visualization of given data

This is our Depended variable, The unemployment rate (UNEMP) which we will model. As one can see, here we have almost linear change of automatically made model.

Hide

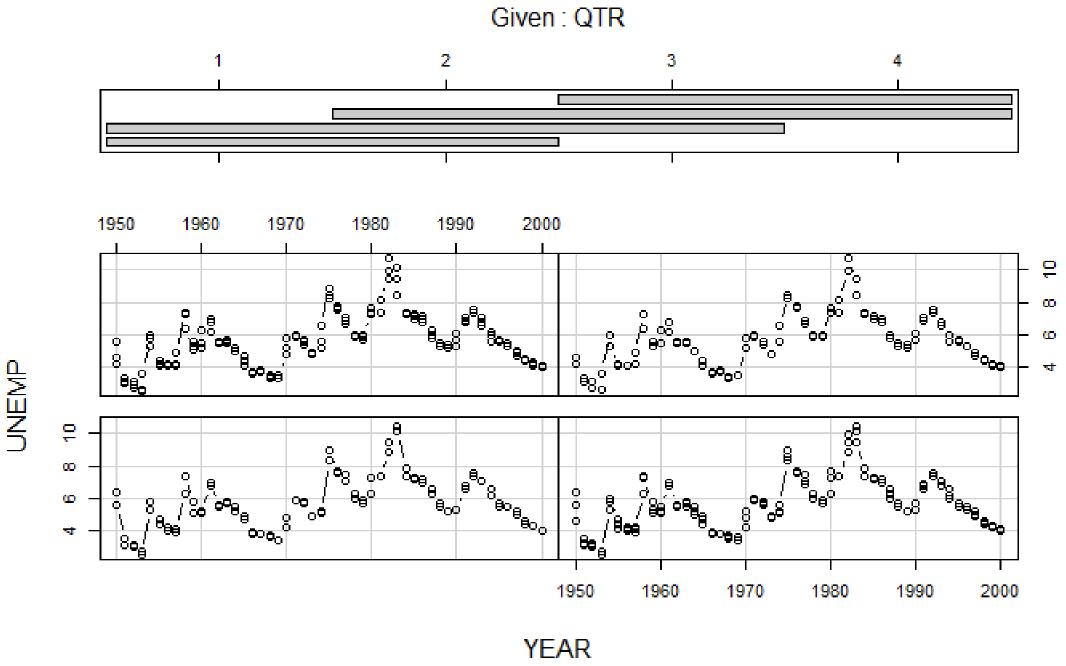
```
scatterplot(UNEMP~YEAR|QTR, boxplots=FALSE, smooth=TRUE, data=UN_data)
```



Here the data shows us the dynamic change of Unemployment rate in the US. This is a little preview of how this rate changes across quarters. Obviously, interwal between quarters must be small, but in further discussion we will test some heterogeneity of them.

Hide

```
coplot(UNEMP ~ YEAR|QTR, type="b", data=UN_data, number = 4)
```



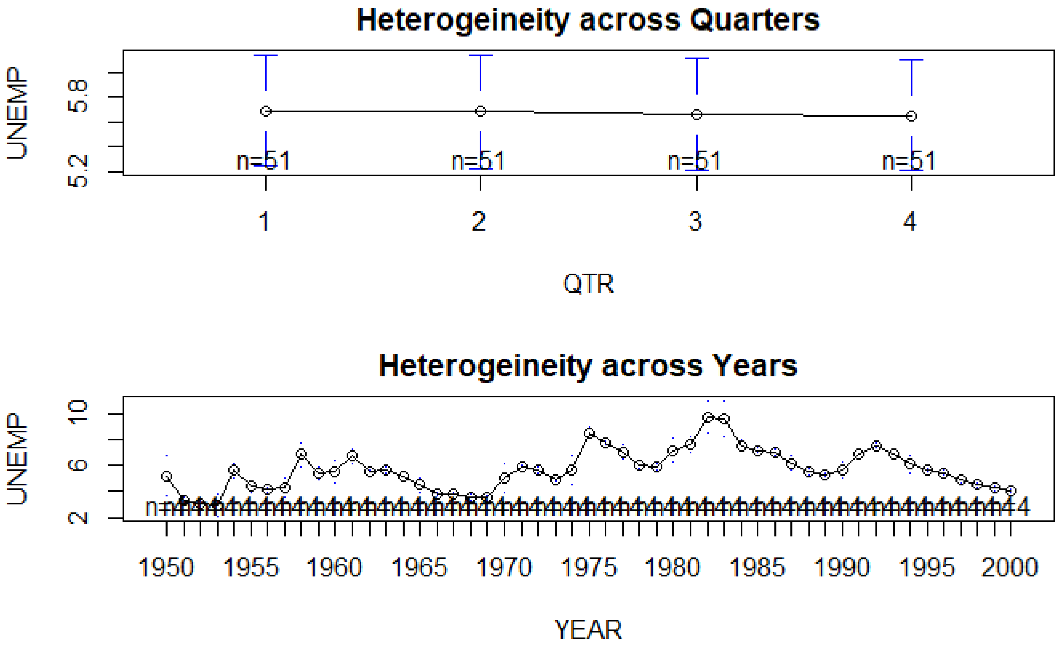
Hide

```
detach("package:foreign")
```

Here, we can see heterogeneity across years and quarters. Even now it seems that unemployment rate is decreasing during the year. Also, heterogeneity during whole time period is dynamically changing.

Hide

```
# plotmeans draw a 95% confidence interval around the means
par(mfrow=c(2,1))
plotmeans(UNEMP ~ QTR, main="Heterogeineity across Quarters", data=UN_data)
plotmeans(UNEMP ~ YEAR, main="Heterogeineity across Years", data=UN_data)
```



Hide

```
detach("package:gplots")
```

Building models

OLS Linear Regression The first one, and most popular way to predict some changes is to build linear regression model, based on Ordinary least squares method of modelling changes. OLS model will allow us to predict the way of Unemployment change. Summary of the model says, all the variables are somehow connected and Adjusted R-squared is very high. Also, residual standard error is small, which is very good. There is also big amount of degrees of freedom.

Hide

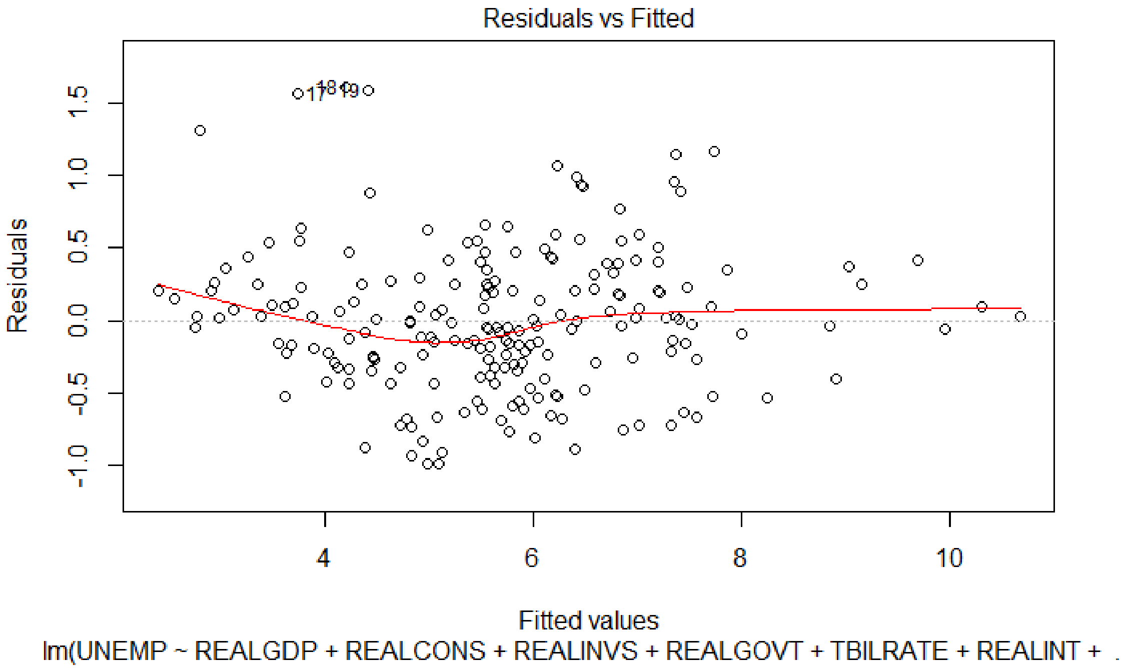
```
ols <- lm(UNEMP~REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE + REALINT + M1 + CPI_U + POP + INFL, dat
a=UN_data)
summary(ols)
```

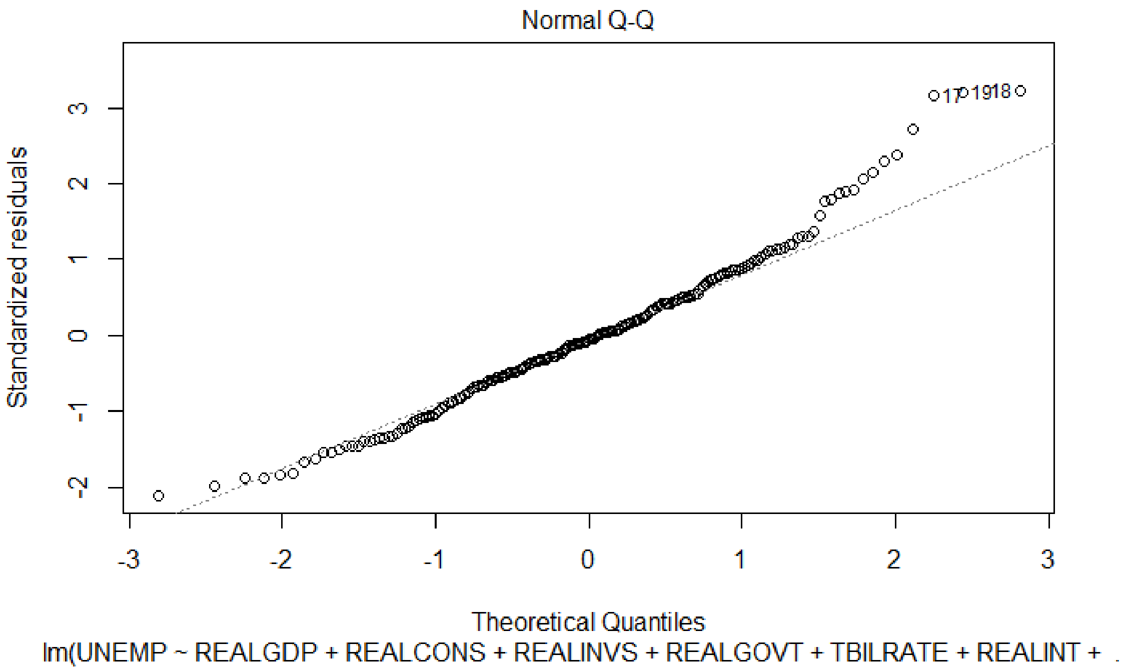
```

Call:
lm(formula = UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT +
    TBILRATE + REALINT + M1 + CPI_U + POP + INFL, data = UN_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98813 -0.31174 -0.02158  0.25227  1.60271

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.861e+01  1.366e+00 -13.627  < 2e-16 ***
REALGDP     -6.574e-03  1.079e-03  -6.090 5.98e-09 ***
REALCONS     6.737e-03  1.217e-03   5.536 1.00e-07 ***
REALINVS    -4.375e-03  1.019e-03  -4.295 2.77e-05 ***
REALGOVT    -9.417e-03  1.082e-03  -8.700 1.45e-15 ***
TBILRATE     8.947e-01  4.801e-01   1.864   0.0639 .
REALINT     -1.064e+00  4.831e-01  -2.202   0.0288 *
M1          -8.835e-03  9.844e-04  -8.975 2.51e-16 ***
CPI_U        3.550e-02  2.784e-03  12.752  < 2e-16 ***
POP          2.021e-01  1.094e-02  18.471  < 2e-16 ***
INFL        -1.077e+00  4.801e-01  -2.243   0.0260 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5042 on 193 degrees of freedom
Multiple R-squared:  0.9026,	Adjusted R-squared:  0.8976
F-statistic: 178.9 on 10 and 193 DF,  p-value: < 2.2e-16
```

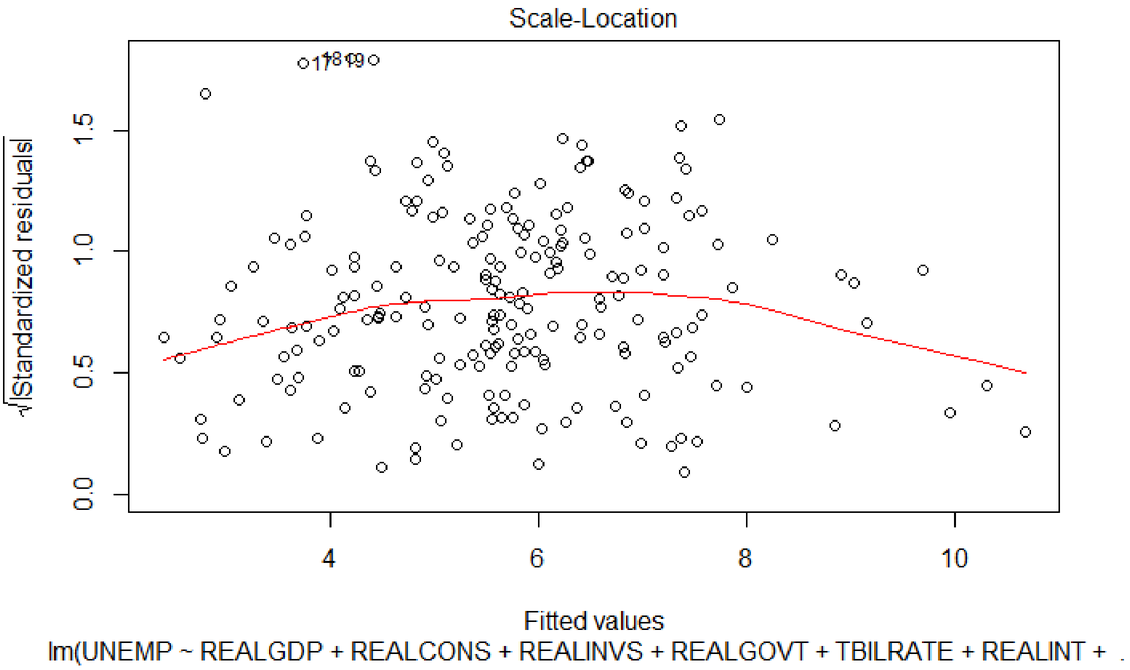
Plotting linear model and checking the correctness, leads us to some observations:
- Residuals vs Fitted plot - line is approximately linear. (which is good)
- Normal Q-Q plot - Distribution is close to Normal. (which is good)
- Scale-Location plot shows that the variance is approximately stable (but a bit increasing)
- Residuals vs Leverage - Our plot doesn’t show any influential cases as all of the cases are within the the dashed Cook’s distance line. - Linear Plot (the last one) shows that variables are changing linearly.

Hide

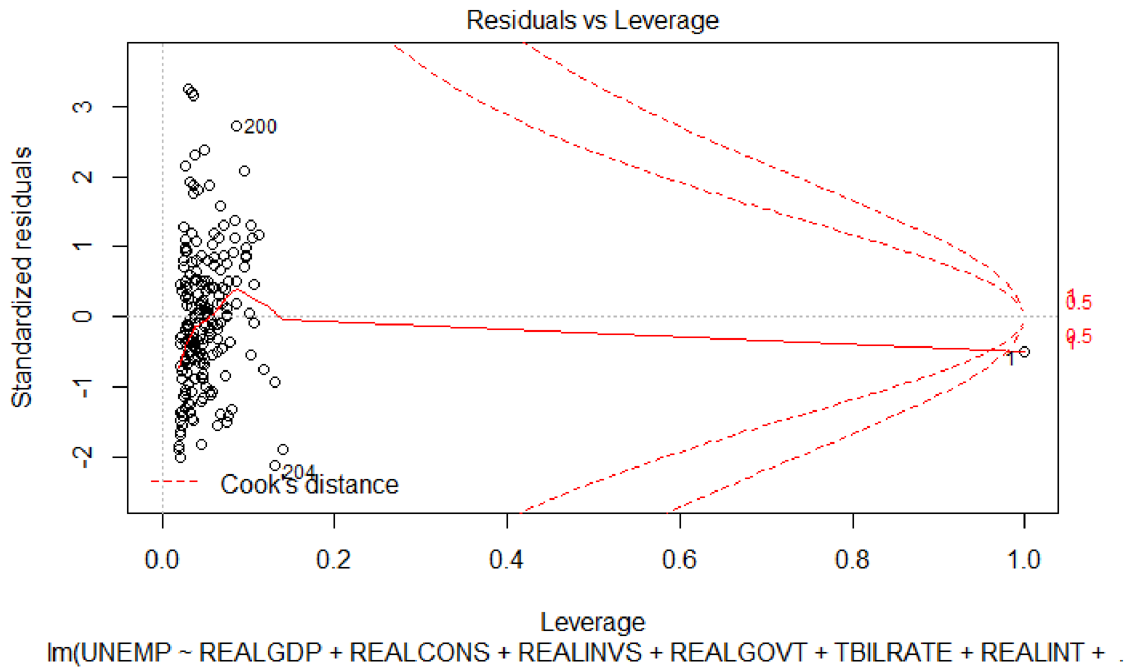
```
plot(ols)
```





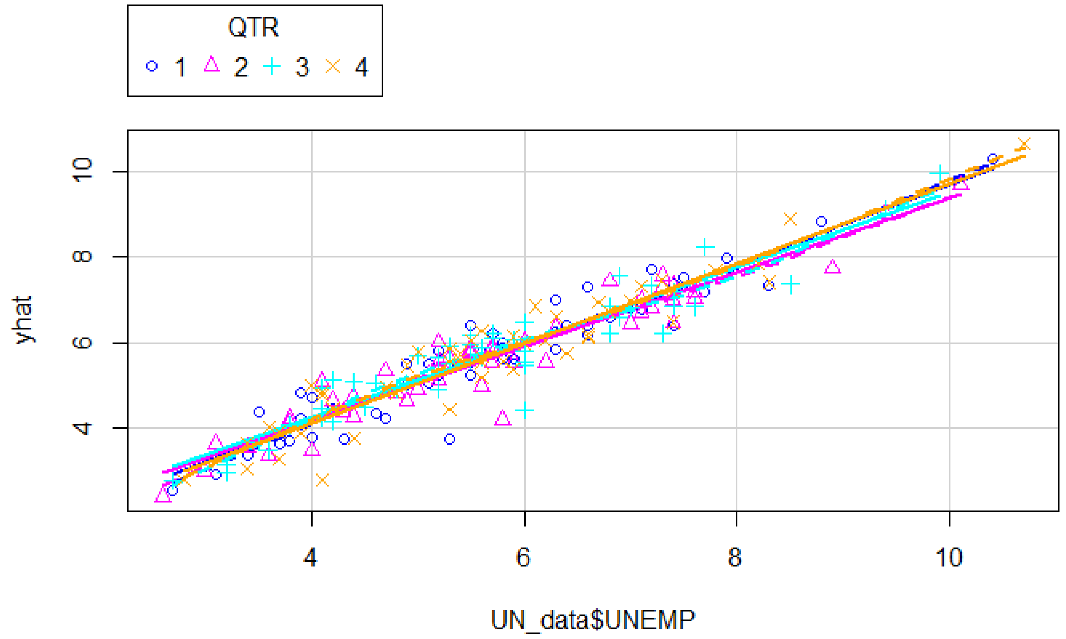


```
NaNs producedNaNs produced
```



Hide

```
yhat <- ols$fitted
scatterplot(yhat~UN_data$UNEMP|QTR, boxplots=FALSE)
```



Modelling Linear fixed effects regression model
As we are trying to test the same rate across different periods of a year - most probably there is little difference, so FE model is prederred. This is our assumption, and we will test it further. For now: Within Estimator model is performed, summary of this model says:
- Now Adjusted R-squared is only ~50%, so it causes us to think of something.
- Real consumption and investment did not played the role, and real GDP has small impact, because we are modelling changes for 4 quarters (and theese variables are left unchanged across one year)
- Also, residual sum of squares got higher.

Hide

```
model.fe <-plm(UNEMP~ QTR +REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE + REALINT + M1 + CPI_U + POP
+ INFL, data = UN_data, model = 'within')
summary(model.fe)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = UNEMP ~ QTR + REALGDP + REALCONS + REALINVS + REALGOVT + 
    TBILRATE + REALINT + M1 + CPI_U + POP + INFL, data = UN_data, 
    model = "within")

Balanced Panel: n = 51, T = 4, N = 204

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max. 
-0.6562417 -0.1220914 -0.0082607  0.1087741  0.6282654 

Coefficients:
            Estimate  Std. Error t-value  Pr(>|t|)    
QTR2      0.10851827  0.06947027  1.5621 0.1205263    
QTR3      0.15885207  0.11128411  1.4274 0.1556781    
QTR4      0.23726606  0.15519506  1.5288 0.1285642    
REALGDP  -0.00466246  0.00183166 -2.5455 0.0119956 *  
REALCONS  0.00056458  0.00172987  0.3264 0.7446306    
REALINVS -0.00010544  0.00201774 -0.0523 0.9583977    
REALGOVT -0.00074643  0.00194203 -0.3844 0.7012994    
TBILRATE  0.99672402  0.25015179  3.9845 0.0001083 ***
REALINT  -1.15057590  0.25278859 -4.5515 1.146e-05 ***
M1       -0.00182785  0.00189254 -0.9658 0.3357989    
CPI_U     0.02545247  0.00869143  2.9285 0.0039778 ** 
POP       0.05156239  0.06498884  0.7934 0.4288853    
INFL     -1.16915902  0.24940663 -4.6878 6.483e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    20.075
Residual Sum of Squares: 7.1333
R-Squared:      0.64467
Adj. R-Squared: 0.48477
F-statistic: 19.5382 on 13 and 140 DF, p-value: < 2.22e-16
```

Modelling Linear random effects regression model.
RE model is necessarily to check, in case we have panel data.
- Adjusted R-squaredgot bigger
- Observing the same proclivity: Consumption, Investment are not important ones.
Further we will perform some diagnostics (including Hausman test for stat. significanse) and test, which model is better.

Hide

```
model.rm <-plm(UNEMP~QTR +REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE + REALINT + M1 + CPI_U + POP + 
INFL, data = UN_data, model = 'random')
summary(model.rm)
```

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = UNEMP ~ QTR + REALGDP + REALCONS + REALINVS + REALGOVT +
    TBILRATE + REALINT + M1 + CPI_U + POP + INFL, data = UN_data,
    model = "random")

Balanced Panel: n = 51, T = 4, N = 204

Effects:
                  var std.dev share
idiosyncratic 0.05095 0.22573 0.203
individual    0.20002 0.44723 0.797
theta: 0.7553

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.590354 -0.164774 -0.045947  0.139426  0.840144

Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -19.3436620   1.9679219 -9.8295 < 2.2e-16 ***
QTR2          0.0305854   0.0514217  0.5948  0.551980
QTR3         -0.0277864   0.0514837 -0.5397  0.589395
QTR4         -0.0212899   0.0538107 -0.3956  0.692367
REALGDP      -0.0042448   0.0013740 -3.0894  0.002005 **
REALCONS      0.0016732   0.0013869  1.2064  0.227659
REALINVS     -0.0021929   0.0014092 -1.5561  0.119675
REALGOVT     -0.0057537   0.0014774 -3.8945 9.838e-05 ***
TBILRATE      1.1139831   0.2772740  4.0176 5.879e-05 ***
REALINT      -1.2629602   0.2799757 -4.5110 6.453e-06 ***
M1           -0.0065547   0.0012967 -5.0550 4.303e-07 ***
CPI_U         0.0341348   0.0036705  9.2998 < 2.2e-16 ***
POP           0.1992193   0.0160076 12.4453 < 2.2e-16 ***
INFL         -1.2753834   0.2766036 -4.6109 4.010e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    49.032
Residual Sum of Squares: 12.423
R-Squared:      0.74665
Adj. R-Squared: 0.72931
Chisq: 559.937 on 13 DF, p-value: < 2.22e-16
```

Diagnostic Tests

Hausman test - The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent. As one can see, the models are statistically significant.

Hide

```
phtest(model.fe, model.rm)
```

```


    Hausman Test

data:  UNEMP ~ QTR + REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE +  ...
chisq = 348.42, df = 13, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

A lagrange Multiplier test is one of three classical approaches to hypothesis testing - model with what effects are better - FE (p-value < 0.05), RE(p-value > 0.05). Here, fixed effects model are preffered.

Hide

```
#Regular OLS (pooling model) using plm
pool <- plm(UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE + REALINT + M1 + CPI_U + POP + INFL,
data = UN_data, model = 'pool')
summary(pool)
```

```
Pooling Model

Call:
plm(formula = UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT +
    TBILRATE + REALINT + M1 + CPI_U + POP + INFL, data = UN_data,
    model = "pool")

Balanced Panel: n = 51, T = 4, N = 204

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.988127 -0.311742 -0.021581  0.252274  1.602710

Coefficients:
               Estimate  Std. Error  t-value  Pr(>|t|)
(Intercept) -1.8611e+01  1.3658e+00 -13.6265 < 2.2e-16 ***
REALGDP     -6.5740e-03  1.0794e-03  -6.0902 5.984e-09 ***
REALCONS     6.7367e-03  1.2170e-03   5.5357 1.003e-07 ***
REALINVS    -4.3747e-03  1.0186e-03  -4.2949 2.766e-05 ***
REALGOVT    -9.4174e-03  1.0824e-03  -8.7001 1.450e-15 ***
TBILRATE     8.9471e-01  4.8011e-01   1.8635   0.06390 .
REALINT     -1.0639e+00  4.8308e-01  -2.2024   0.02882 *
M1          -8.8353e-03  9.8439e-04  -8.9754 2.511e-16 ***
CPI_U        3.5501e-02  2.7841e-03  12.7516 < 2.2e-16 ***
POP          2.0209e-01  1.0941e-02  18.4715 < 2.2e-16 ***
INFL        -1.0770e+00  4.8013e-01  -2.2431   0.02603 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    503.73
Residual Sum of Squares: 49.056
R-Squared:      0.90261
Adj. R-Squared: 0.89757
F-statistic: 178.879 on 10 and 193 DF, p-value: < 2.22e-16
```

Hide

```
plmtest(pool)
```

```

    Lagrange Multiplier Test - (Honda) for balanced panels

data:  UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE +  ...
normal = 10.87, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Testing for cross-sectional dependence/contemporaneous correlation: using Pasaran CD test and Testing for serial correlation As one can see,there are no serial correlation, p-value = 0.0118 but there is some cross-sectional dependence between variables.

Hide

```
pcdtest(model.fe, test = c("cd"))
```

```

    Pesaran CD test for cross-sectional dependence in panels

data:  UNEMP ~ QTR + REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE +     REALINT + M1 + CPI_U + POP + 
INFL
z = -0.93039, p-value = 0.3522
alternative hypothesis: cross-sectional dependence
```

Hide

```
pbgtest(model.fe)
```

```

    Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  UNEMP ~ QTR + REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE +     REALINT + M1 + CPI_U + POP +
INFL
chisq = 30.358, df = 4, p-value = 4.139e-06
alternative hypothesis: serial correlation in idiosyncratic errors
```

Testing for unit roots/stationarity Dickey-Fuller Test

Hide

```
Panel.set <- plm.data(UN_data)
```

```
use of 'plm.data' is discouraged, better use 'pdata.frame' instead
```

Hide

```
adf.test(Panel.set$UNEMP, k=2)
```

```

    Augmented Dickey-Fuller Test

data:  Panel.set$UNEMP
Dickey-Fuller = -2.9584, Lag order = 2, p-value = 0.1745
alternative hypothesis: stationary
```

The null hypothesis for the Breusch-Pagan test is homoskedasticity. Result: homoskedasticity is in this data.

Hide

```
bptest(UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE + REALINT + M1 + CPI_U + POP + INFL, data
=UN_data, studentize=F)
```

```

    Breusch-Pagan test

data:  UNEMP ~ REALGDP + REALCONS + REALINVS + REALGOVT + TBILRATE +     REALINT + M1 + CPI_U + POP + INFL
BP = 30.385, df = 10, p-value = 0.0007408
```

Part 2 ##Hypothesis Testing Let`s perform t-tests on our models

Hide

```
print("-----Linear Ordinary least Squares Model------")
```

```
[1] "-----Linear Ordinary least Squares Model------"
```

Hide

```
coeftest(ols, vcov. = vcovHC, type = "HC1")
```

```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept) -1.8611e+01  1.2983e+00 -14.3350 < 2.2e-16 ***
REALGDP     -6.5740e-03  9.3142e-04  -7.0580 2.961e-11 ***
REALCONS     6.7367e-03  1.0926e-03   6.1659 4.020e-09 ***
REALINVS    -4.3747e-03  9.9494e-04  -4.3970 1.811e-05 ***
REALGOVT    -9.4174e-03  9.0143e-04 -10.4472 < 2.2e-16 ***
TBILRATE     8.9471e-01  1.8489e-01   4.8392 2.662e-06 ***
REALINT     -1.0639e+00  1.8981e-01  -5.6050 7.117e-08 ***
M1          -8.8353e-03  9.1697e-04  -9.6353 < 2.2e-16 ***
CPI_U        3.5501e-02  2.2931e-03  15.4813 < 2.2e-16 ***
POP          2.0209e-01  1.0158e-02  19.8939 < 2.2e-16 ***
INFL        -1.0770e+00  1.8366e-01  -5.8640 1.930e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
print("-----Linear Model with Fixed effects-----")
```

```
[1] "-----Linear Model with Fixed effects-----"
```

Hide

```
coeftest(model.fe, vcov. = vcovHC, type = "HC1")
```

```
t test of coefficients:

            Estimate  Std. Error t value  Pr(>|t|)    
QTR2      0.10851827  0.06767500  1.6035   0.11107    
QTR3      0.15885207  0.09172883  1.7318   0.08552 .  
QTR4      0.23726606  0.12199755  1.9448   0.05380 .  
REALGDP  -0.00466246  0.00236382 -1.9724   0.05053 .  
REALCONS  0.00056458  0.00214779  0.2629   0.79304    
REALINVS -0.00010544  0.00276783 -0.0381   0.96967    
REALGOVT -0.00074643  0.00161618 -0.4618   0.64491    
TBILRATE  0.99672402  0.11637111  8.5650 1.759e-14 ***
REALINT  -1.15057590  0.13348538 -8.6195 1.290e-14 ***
M1       -0.00182785  0.00141109 -1.2953   0.19733    
CPI_U     0.02545247  0.00890476  2.8583   0.00491 ** 
POP       0.05156239  0.04545837  1.1343   0.25862    
INFL     -1.16915902  0.11975389 -9.7630 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
print("-----Linear Model with Random effects-----")
```

```
[1] "-----Linear Model with Random effects-----"
```

Hide

```
coeftest(model.rm, vcov. = vcovHC, type = "HC1")
```

```

t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept) -19.3436620   2.3276546  -8.3104 1.783e-14 ***
QTR2          0.0305854   0.0461650   0.6625  0.508438    
QTR3         -0.0277864   0.0476690  -0.5829  0.560649    
QTR4         -0.0212899   0.0595588  -0.3575  0.721144    
REALGDP      -0.0042448   0.0015174  -2.7975  0.005681 ** 
REALCONS      0.0016732   0.0015374   1.0883  0.277839    
REALINVS     -0.0021929   0.0016433  -1.3345  0.183638    
REALGOVT     -0.0057537   0.0019216  -2.9942  0.003118 ** 
TBILRATE      1.1139831   0.1193008   9.3376 < 2.2e-16 ***
REALINT      -1.2629602   0.1267539  -9.9639 < 2.2e-16 ***
M1           -0.0065547   0.0012352  -5.3067 3.091e-07 ***
CPI_U         0.0341348   0.0037858   9.0167 < 2.2e-16 ***
POP           0.1992193   0.0183059  10.8828 < 2.2e-16 ***
INFL         -1.2753834   0.1157596 -11.0175 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Last question

Is it true, that in average, the biggest unemployment rate is at the beggining of the year?

Hide

```
Q1_data = UN_data[which(UN_data$QTR == 1), names(UN_data) %in% c("UNEMP")]
Q2_data = UN_data[which(UN_data$QTR == 2), names(UN_data) %in% c("UNEMP")]
Q3_data = UN_data[which(UN_data$QTR == 3), names(UN_data) %in% c("UNEMP")]
Q4_data = UN_data[which(UN_data$QTR == 4), names(UN_data) %in% c("UNEMP")]
c(mean(Q1_data), "- First Quarter")
```

```
[1] "5.69411764705882" "- First Quarter"
```

Hide

```
c(mean(Q2_data), "- Second Quarter")
```

```
[1] "5.68627450980392" "- Second Quarter"
```

Hide

```
c(mean(Q3_data), "- Third Quarter")
```

```
[1] "5.66274509803922" "- Third Quarter"
```

Hide

```
c(mean(Q4_data), "- Fourth Quarter")
```

```
[1] "5.65490196078431" "- Fourth Quarter"
```

Hide

```
c(max(c(mean(Q1_data), "- First Quarter"),c(mean(Q2_data), "- Second Quarter"),c(mean(Q3_data), "- Third Qua
rter"),c(mean(Q4_data), "- Fourth Quarter")), "Maximum")
```

```
[1] "5.69411764705882" "Maximum"
```

Hide

```
if (c(max(c(mean(Q1_data), "- First Quarter"),c(mean(Q2_data), "- Second Quarter"),c(mean(Q3_data), "- Third
Quarter"),c(mean(Q4_data), "- Fourth Quarter")), "Maximum")
==c(mean(Q1_data), "- First Quarter")){
  print("Null hypothesis accepted")}else{ print("Null Hypothesis rejected")}
```

```
the condition has length > 1 and only the first element will be used
```

```
[1] "Null hypothesis accepted"
```

Conclusions

This project is forced to show the linear dependence of unemplyment rate, Real GDP in the US dollars, Real disposable personal income, Consumer price index, Nominal money stock, Quarterly average of month end 90 day t bill rate, Pop = Population, Rate of inflation, and real interest rate and to get models for predicting Unemployment rates in next years. Also, we saw that the biggest unemployment rate is, in average at the beginning of the year. This project show the power of regression analysis andthe signifficanse of it`s usage.

References

Data source http://people.stern.nyu.edu/wgreene/Text/Edition7/TableF5-2.txt
Understanding Panel data Regression https://towardsdatascience.com/understanding-panel-data-regression-c24cd6c5151e