Midterm project

Congyao Duan

2020/11/13

1.Abstract

Economic freedom is vital to a country's consumption ability,so I think the report can help companies to decide which country's market is more valuable to exploit. I use linear regression model,multilevel regression model and time series model to analyze the changes of economic freedom and the factors that affect economic freedom in this report.

2.Instruction

The project I plan to do is about economic freedom.What I 'm going to do is Analyzing the factors that affect the economic freedom& Analyzing the changing trend and predict the economic freedom index for different countries in the future.

I did some research about economic freedom and economic development of one country.I found out that the factors that affact economic development and economic freedom in a country are economic development, higher education, natural factors and people's income.So I collected these datasets to analyze their impact on economic freedom

3.Data Cleaning and Praparing

3.1 Data source

1. https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=1&year=2017 (https://www.fraserinstitute.org/economic-freedom/dataset?geozone=world&page=dataset&min-year=2&max-year=0&filter=1&year=2017) This web page provides the information about the economic freedom in different countries from 1970-2018. It also provide the size of government, legal system& property rights, sound money, freedom to trade internationally, regulation(the factors that affect the economic freedom).
2. https://github.com/datasets/gdp/blob/master/data/gdp.csv (https://github.com/datasets/gdp/blob/master/data/gdp.csv) GDP for different countries in different years
3. https://github.com/datasets/population/blob/master/data/population.csv (https://github.com/datasets/population/blob/master/data/population.csv) Population for different countries in different years
4. https://github.com/datasets/cpi/blob/master/data/cpi.csv (https://github.com/datasets/cpi/blob/master/data/cpi.csv) Customer price index for different countries in different years
5. https://github.com/datasets/expenditure-on-research-and-development (https://github.com/datasets/expenditure-on-research-and-development) Higher education and research funding

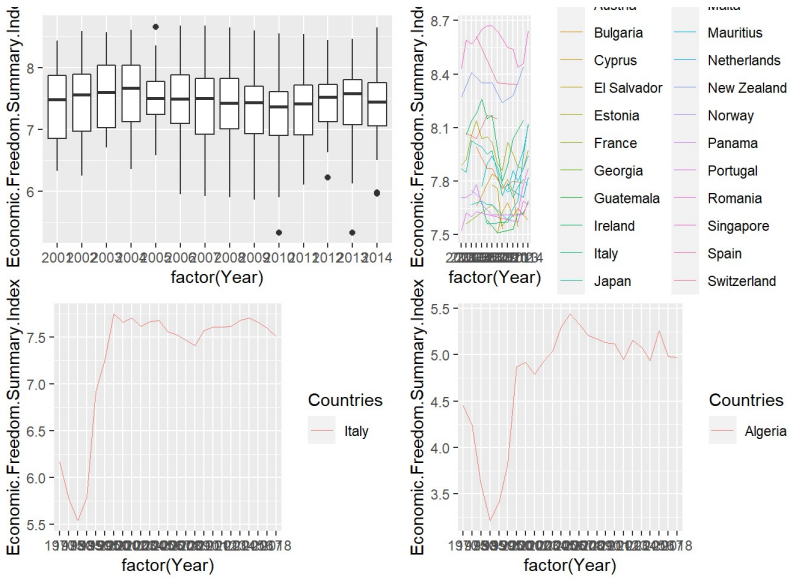
3.2 Data cleaning and orgnizing

In this part, I made two data sets.I add the data I need to the dataset ,filter the required columns separately, remove the unneeded data, and remove the NA.(dataset eco is to analysis the Index of economic freedom in different years and dataset eco2 is to analysis the factors that can affect economic freedom)

4 EDA

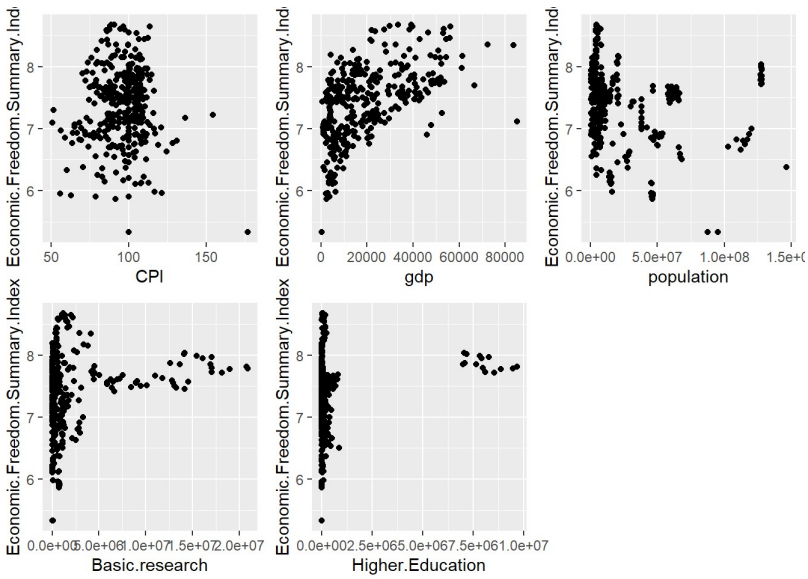
4.1 Index of economic freedom in different years

To study trends in economic freedom, I drew boxplots and line charts (too many country lines are too confusing, so I divided them into four plots to draw all the countries,one of the four plots is shown in the report). Then I selected two countries with a large gap and I was more interested in to make separate line charts. From the EDA part,I find out that economic freedom in the world as a whole is on the rise, but the index of economic freedom in individual countries is not always on the rise.



4.2 Factors that can affect economic freedom

I drew a scatter plot of economic freedom and the different factors that influence it. It can be seen that these factors have a certain impact on economic freedom.



5 Models and Interpretation&Validation

5.1 Index of economic freedom in different years

I start with the linear regression.But linear models are not necessarily appropriate. The model fitted with the economic free values of all countries must have a large error in the prediction.Then I select 12 countries to fit multilevel model.

I also did some research on how people deal with this kind of problem.I found out that time series model may be a good choice,so I read a book and some of articles about time series and also try this model on individual country.I chose one of the higher-ranked countries and one of the lower-ranked countries to fit the model.

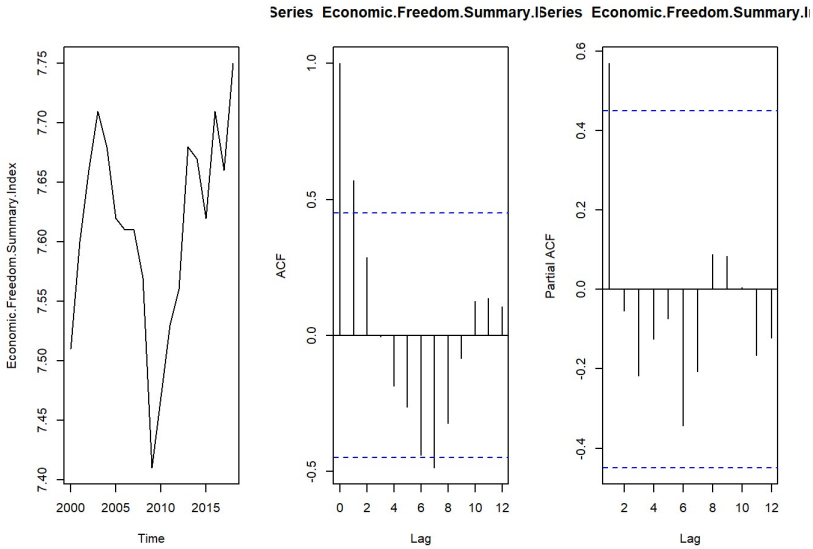
(Time-series model: time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time Series analysis can be useful to see how a given asset, security or economic variable changes over time.)

I fit the time series model in this order:Verify that if the dataset is a stationary white noise sequence,calculate ACF/PACF,ARIMA model recognition,model check,prediction.

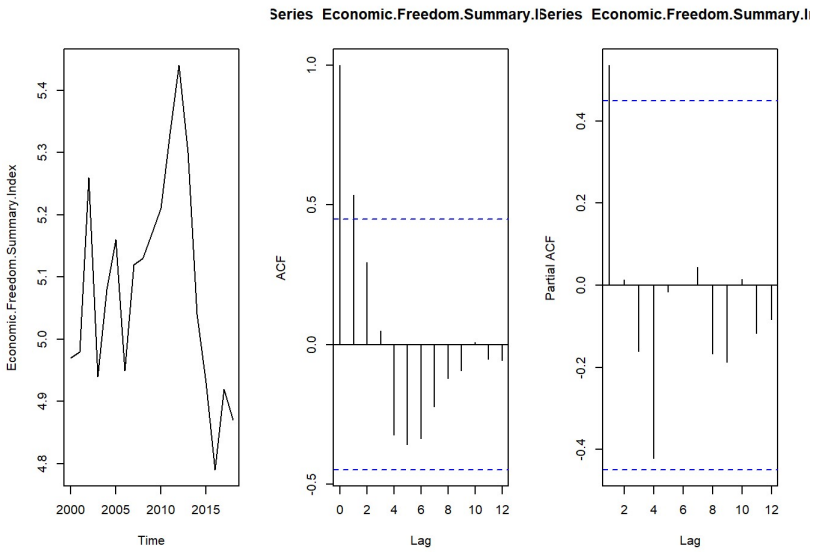
1 Fit models

By fitting the linear model, I can see that the world as a whole has an exponential increase in economic freedom over time(after 1995).But I don't think it's a good model for predicting economic freedom in different countries.So I try to fit multilevel model.

I also fit time series models for individual countries.Italy is the country with the high index of economic freedom.By the white noise sequence test I see that p-value<0.05,so this dataset is not a white noise sequence.Then I observed whether the AIC/PAIC diagram was towed or truncated, and use system fomula(auto.arima) determined appropriate ARIMA model.It shows that AR(1) model is suitable for this two dataset.

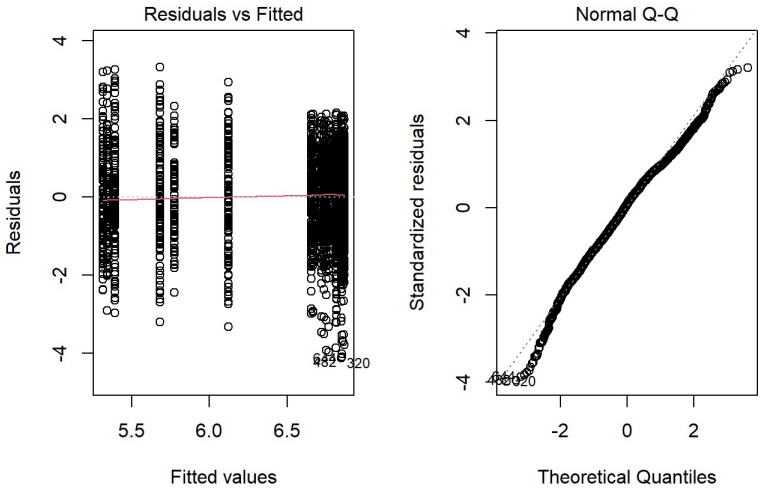


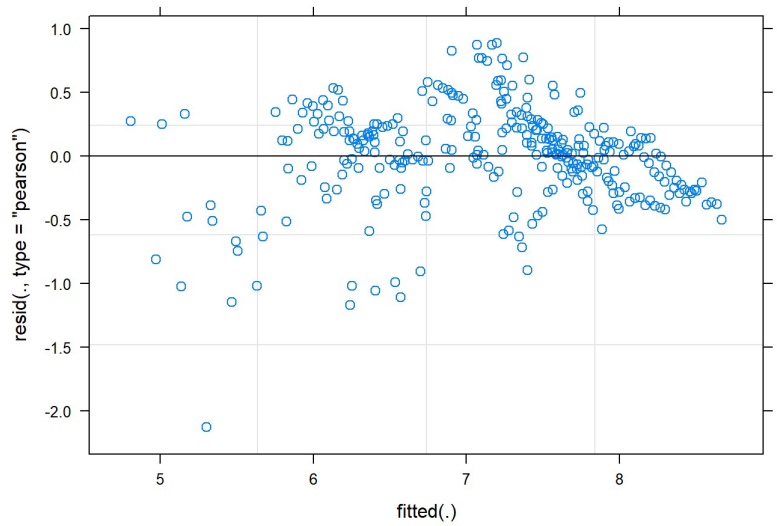
Algeria is the country with the low index of economic freedom.By the white noise sequence test I see that p-value<0.05,so this dataset is not a white noise sequence.Then I observed whether the AIC/PAIC diagram was towed or truncated, and use system fomula(auto.arima) determined appropriate ARIMA model.It shows that AR(1) model is suitable for this dataset.



2 Model Validation

Draw residual plot and QQ plot of linear regression model and multilevel model.As I suspected,The residual plot of linear regression model looks bad(the range of residuals is too large). And R^2of this model is very low.As for multilevel regression model,the coefficient is very small.And as we said before,changes in the index of economic freedom do not show particular trends, so I think that multilevel regression is not a good choice,either.





The test method of time series model is to test whether the residual is white noise series.If the residual is a white noise sequence, it means the model does not omit useful information, so the model is appropriate. By checking the two time series models,I found that all the p-values are larger than 0.05,it means that I cannot reject the null hypothesis(the residuals are listed as white noise sequences).So the model is significantly effective.

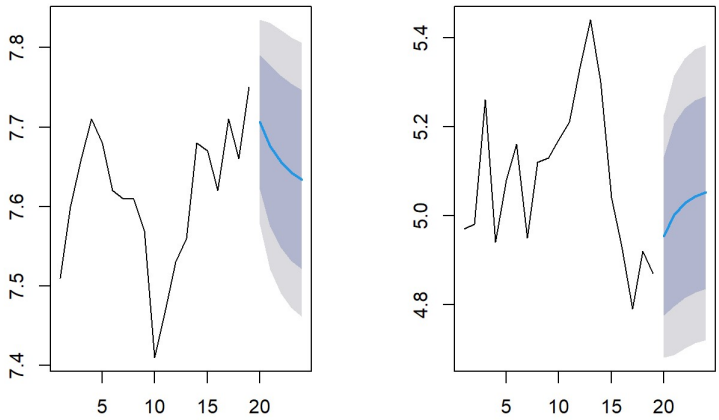
3 Prediction

The fitted model is used to predict the economic freedom index of these two countries(Italy&Algeria) in the next five years:

```
##    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 20       7.706119 7.622100 7.790139 7.577623 7.834616
## 21       7.676534 7.575202 7.777867 7.521559 7.831509
## 22       7.656587 7.548295 7.764878 7.490969 7.822204
## 23       7.643138 7.531827 7.754449 7.472902 7.813374
## 24       7.634070 7.521413 7.746727 7.461776 7.806364
```

```
##    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 20       4.953130 4.774583 5.131676 4.680066 5.226193
## 21       5.000564 4.794995 5.206133 4.686174 5.314954
## 22       5.027630 4.814000 5.241261 4.700911 5.354350
## 23       5.043074 4.826884 5.259265 4.712440 5.373709
## 24       5.051887 4.834870 5.268904 4.719988 5.383786
```

ecasts from ARIMA(1,0,0) with non-ze **ecasts from ARIMA(1,0,0) with non-ze**



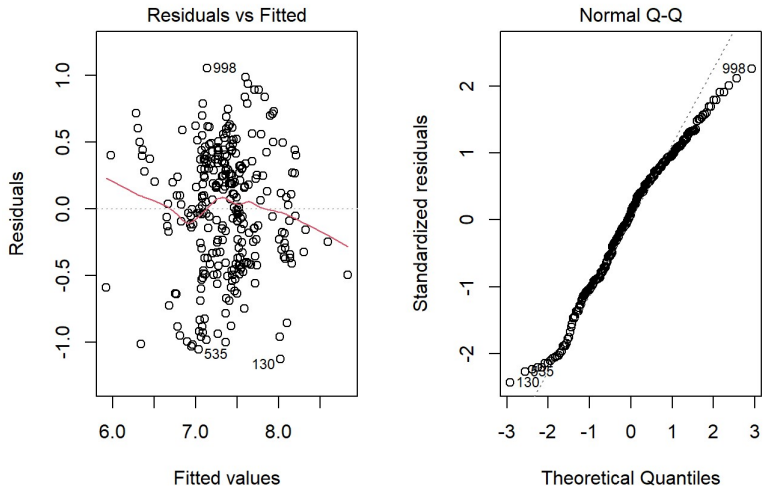
5.2 Factors that can affect economic freedom

1 Fit models

In this part,I divided the data set into two parts: the training set and the test set(80 percent and 20 percent, respectively).By observing the data, I think this data set is suitable for fitting with a linear model.

2 Model Validation

The R^2 of this model is ok and p-value is much smaller than 0.05.All the independent variables are significant except Basic research(which is 0.23,I think it is acceptable). I use cross validation,residual plot and QQ plot to validate this model.By cross validation I get the value of RMSE=0.46 and MAE=0.38.Residual plot looks ok and qq plot looks ok too.



6 Discussion

According to all the EDA and model above, we can conclude that,with the change of year, the change of economic freedom index is not monotonically increasing or decreasing(for one country).Linear regression model and multilevel model cannot explain the change in the data very well.Time series model can explain the data much better and I predict the future direction of economic freedom index.As for factors that affect countries' economy index,I think the GDP,CPI,population,higher education and research funding can affect counties' economic freedom.

7 Reference

https://www.google.com/search?source=hp&ei=fwPRX67fO4ODr7wP_a2S6A8&q=the+factors+that+affect+countires%27+economy&oq=the+factors+that+affect+countires%27+economy&gs_lcpab&ved=0ahUKEwjump7dsMHtAhWDwYsBHf2WBP0Q4dUDCAY&uact=5 (https://www.google.com/search?source=hp&ei=fwPRX67fO4ODr7wP_a2S6A8&q=the+factors+that+affect+countires%27+economy&oq=the+factors+that+affect+countires%27+economy&gs_lcpab&ved=0ahUKEwjump7dsMHtAhWDwYsBHf2WBP0Q4dUDCAY&uact=5)

时间序列分析——基于R(Time series analysis–R) 中国人民大学出版社（China Renmin University Press）

https://www.google.com/search?source=hp&ei=_-7NX5-HF8H4hwOBpIrIBQ&q=time+series+model&oq=time+series+model&gs_lcp=CgZwc3ktYWIQAzICCAAyAggAMgIIADICCAAyAggAMgIIADICab&ved=0ahUKEwjfgsHgwLvtAhVB_GEKHQGSAlkQ4dUDCAY&uact=5 (https://www.google.com/search?source=hp&ei=_-7NX5-HF8H4hwOBpIrIBQ&q=time+series+model&oq=time+series+model&gs_lcp=CgZwc3ktYWIQAzICCAAyAggAMgIIADICCAAyAggAMgIIADICab&ved=0ahUKEwjfgsHgwLvtAhVB_GEKHQGSAlkQ4dUDCAY&uact=5)

Appendix

model1:lm(Economic.Freedom.Summary.Index~factor(Year),data=eco2)

```
##
## Call:
## lm(formula = Economic.Freedom.Summary.Index ~ factor(Year), data = eco2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1236 -0.6987  0.0704  0.7598  3.3185
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.77417    0.11325  50.987  < 2e-16 ***
## factor(Year)1975 -0.42888    0.15162  -2.829  0.00470 **
## factor(Year)1980 -0.45750    0.15010  -3.048  0.00232 **
## factor(Year)1985 -0.37903    0.14953  -2.535  0.01129 *
## factor(Year)1990 -0.09263    0.14843  -0.624  0.53265
## factor(Year)1995  0.34567    0.14691   2.353  0.01869 *
## factor(Year)2000  0.87807    0.14644   5.996 2.23e-09 ***
## factor(Year)2001  0.89030    0.14691   6.060 1.51e-09 ***
## factor(Year)2002  0.94437    0.14691   6.428 1.47e-10 ***
## factor(Year)2003  0.95898    0.14597   6.570 5.82e-11 ***
## factor(Year)2004  0.98876    0.14530   6.805 1.19e-11 ***
## factor(Year)2005  0.97520    0.14306   6.817 1.10e-11 ***
## factor(Year)2006  1.03456    0.14306   7.232 5.87e-13 ***
## factor(Year)2007  1.07541    0.14306   7.517 7.12e-14 ***
## factor(Year)2008  1.03867    0.14306   7.260 4.76e-13 ***
## factor(Year)2009  1.04179    0.14306   7.282 4.06e-13 ***
## factor(Year)2010  1.03884    0.14095   7.370 2.13e-13 ***
## factor(Year)2011  1.07106    0.14095   7.599 3.84e-14 ***
## factor(Year)2012  1.07002    0.14095   7.592 4.06e-14 ***
## factor(Year)2013  1.08386    0.14031   7.725 1.47e-14 ***
## factor(Year)2014  1.08948    0.14000   7.782 9.42e-15 ***
## factor(Year)2015  1.07602    0.14000   7.686 1.98e-14 ***
## factor(Year)2016  1.07664    0.13955   7.715 1.58e-14 ***
## factor(Year)2017  1.06948    0.13955   7.664 2.35e-14 ***
## factor(Year)2018  1.08608    0.13955   7.783 9.37e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 3382 degrees of freedom
## Multiple R-squared:  0.1842, Adjusted R-squared:  0.1785
## F-statistic: 31.83 on 24 and 3382 DF,  p-value: < 2.2e-16
```

model2:lmer(Economic.Freedom.Summary.Index~Year+(1|Countries),data=eco3)

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Economic.Freedom.Summary.Index ~ Year + (1 | Countries)
##    Data: eco3
##
## REML criterion at convergence: 365.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3280 -0.5341  0.0797  0.5680  2.2253
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Countries (Intercept) 0.5559   0.7456
##  Residual              0.1599   0.3999
## Number of obs: 300, groups:  Countries, 12
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -59.119302   3.569991  -16.56
## Year          0.033087   0.001779   18.59
##
## Correlation of Fixed Effects:
##      (Intr)
## Year -0.998
```

model3:arima(Ita$Economic.Freedom.Summary.Index,order=c(1,0,0))

```
##
## Call:
## arima(x = Ita$Economic.Freedom.Summary.Index, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.6742     7.6153
## s.e.  0.1822     0.0419
##
## sigma^2 estimated as 0.004298:  log likelihood = 24.51,  aic = -43.02
##
## Training set error measures:
##                       ME       RMSE        MAE        MPE      MAPE      MASE
## Training set 0.005186949 0.06556076 0.04745981 0.06058132 0.6242105 0.8059213
##                    ACF1
## Training set 0.06231578
```

model4:arima(alg$Economic.Freedom.Summary.Index,order=c(1,0,0))

```
##
## Call:
## arima(x = alg$Economic.Freedom.Summary.Index, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.5706     5.0636
## s.e.  0.1894     0.0710
##
## sigma^2 estimated as 0.01941:  log likelihood = 10.29,  aic = -14.58
##
## Training set error measures:
##                       ME      RMSE       MAE         MPE     MAPE      MASE
## Training set 0.003691605 0.1393208 0.1193732 -0.00329679 2.346614 0.9104738
##                      ACF1
## Training set -0.005986732
```

model5:lm
(Economic.Freedom.Summary.Index~Higher.Education+Basic.research+population+gdp+CPI

```
##
## Call:
## lm(formula = Economic.Freedom.Summary.Index ~ Higher.Education +
##     Basic.research + population + gdp + CPI, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1267 -0.3558  0.0422  0.3543  1.0523
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.582e+00  1.883e-01  40.270  < 2e-16 ***
## Higher.Education  9.010e-08  3.376e-08   2.669  0.00805 **
## Basic.research    2.000e-08  1.672e-08   1.196  0.23277
## population       -8.923e-09  1.329e-09  -6.713 1.05e-10 ***
## gdp               2.026e-05  2.216e-06   9.141  < 2e-16 ***
## CPI              -4.674e-03  1.991e-03  -2.348  0.01955 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4677 on 282 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4451
## F-statistic: 47.04 on 5 and 282 DF,  p-value: < 2.2e-16
```