CPI Analysis

Zachary Palmore

5/16/2021

Abstract

Key Words

Introduction

There are a plethora of entities dedicated, at least in part, to interpreting or predicting the Consumer Price Index (CPI) and rightfully so. An accurate understanding of these statistics offers private and public businesses and governing bodies alike the opportunity to make better fiscal decisions such as when and how much to adjust interest rates, allocate funds, control investments, and raise wages to name a few. Inflation, another highly valued metric for all lending institutions and consumers with credit investments, is found directly from the change in CPI further increasing its value. As stated in the Journal of Economic Perspectives, "Accurately measuring prices and their rate of change, inflation, is central to almost every economic issue."

The Economic Research Service of the United States Department of Agriculture and National Academy of Sciences have written since at least 2006 that reevaluating how the government measures poverty using a new cost of living estimate that utilizes aspects of CPI would elucidate the true nature of impoverished areas and would change who our algorithms consider poor. This study highlights the necessity for new measures in identifying groups of people who need the aid most which thereby reduces currently undetected systemic injustices, reduces waste, and helps to eliminate poverty. In short, use of the CPI in conjunction with other metrics would make each dollar spent on programs like medicare, medicaid, and other welfare assistance programs stretch farther.

Due to the Bureau of Labor Statistics collection of consumer prices, officials can accurately estimate the price of goods and in some regards, it may also serve as a reasonable estimation of the cost of living for individuals. Although, caution should be heeded for those who use the CPI in this manner as its functionality plummets when applied broadly and vastly as a discrete inflation measure. Of course, this is entirely reliant on the validity of the CPI.

One study by an Advisory committee established by the Senate Finance Committee pursuant to a Senate Resolution and documented in the Journal of Economic Perspectives, found that "Over a dozen years, the cumulative additional national debt from overindexing the budget would amount to more than $1 trillion." In the same paper, they found the true CPI value was only 1.1 percentage points over on average per year with a range of 0.8 to 1.6 average percentage points per year. A minor difference that in essence meant a CPI reported to be 3 percent, was actually closer to 2 percent. However, this is not to say that the CPI is any less valuable.

When measuring geographic variations in the cost of living to cross-examine with the academic performance of children, the effects of CPI when used as an indicator for the cost of living (COL) have suggested higher costs were associated with lower achievement. Simultaneously, investments in resources for those same children resulted in better outcomes (often through academic scores) for children in lower-income families. This relationship between the CPI and COL thus has significant implications for assessing and improving

the academic performance of those in lower-income brackets, but the reasons to understand and predict CPI do not stop there.

Regular updates of the CPI offers invaluable information that is central to controlling for economic stability and enacting policies that benefit society. Whether the entities are considering interest rates, monetary accountability, or even education, the CPI has an integral role to play and its accuracy is imperative to support wise decision making. For these reasons, we attempt to build a model that captures the bulk of information crucial in these efforts with a different approach during model development.

Literature Review

Previous research has shown that difficulties in modeling the CPI mostly stem from the inherently unpredictable nature of inflation and consumer prices. This phenomenon is also known as volatility. The agglomeration of changes in prices of consumer goods and services across the US contains an immense diversity of factors influencing those prices. Consider, for example, the cost of water utilities between desert climates such as in Arizona, New Mexico, and Nevada, with the relatively water rich humid subtropical areas of Mississippi, Alabama, and Lousianna. Holding relative demand constant, the cost in utilities to collect or extract, maintain and deliver the supply of this resource will vary considerably and thus, their prices fluctuate according to largely separate variables. Keep in mind this is merely one fluctuating consumer good in a basket of over 80,000 items.

According to the Bureau of Labor Statistics and several other nongovernmental agencies, over the lifetime of the Federal Reserve System (Fed) there has been a concerted effort to keep the inflation rate low. Recall that this inflation rate is determined by finding the change in CPI over time and recorded as an increase or decrease in each calculation based only on the previous value. In this endeavour, the system tends to focus on nonvolatile components which they call the core inflation rate. Studies based on this core inflation rate (which generally excludes food and energy utilities, though not exclusively) have shown that modeling both goods and services "have not fared well in general." One justifiable reason that aggregation of this kind tends to reflect poorly on model testing is that the core inflation measure used by the Fed does not downselect the consumer goods or services well enough to observe separate inflationary expectations. Doing so, may also result in another conundrum, that further reduction is necessary until isolated groups all share distinctly similar volatility in their CPI. Thus, an alternative modeling technique has been proposed.

As published in the Journal of Current Issues in Economics and Finance, several researchers split the data into two groups, one that models the CPI for goods, and one model for services with a critical success; more accurate predictions. With this novel modeling process they captured reality better than traditional core inflation models and were able to discover the behavior of each group and its dependencies in more detail. The observed CPI of goods depended on short-term expectations and import prices while services depended on long-term expectation and leeway of the domestic labor market. Moreover, composite CPI models like this improved the prediction without the use of additional external factors when measured through inflationary reactions.

Today, the Phillips Inflation Curve Forecast remains the most useful macroeconomic model when interpreting or predicting inflation with this univariate model outperforming multivariate counterparts. However, their performance is episodic, limiting its ability to perform well continuously. Some also refer to these more novel models as New Keynesian Phillips Curves (NKPC) models and broadly speaking they employ the use of an activity variable, which often simply means they consider historical CPI values.

Evidence to characterize the backward-looking Phillips Inflation Curve has shown that hawkish monetary policy approaches tend to " trigger an endogenous anchoring of agents' subjective inflation forecasts" and offer "a coherent explanation for all of the observed changes in U.S. inflation behavior." The literature also indicates that volatility in CPI has decreased since the 1980's along with the persistence of US inflation to increase. Those familiar with the relation of inflation rates with the output gap have noted that stability in the Philips Inflation Curve Forecasts have resurfaced from U.S. data since 1980. Additionally, studies done with this backward-looking Phillips model that improve their choice of anchoring (or benchmark) years have

been shown to produce statistically significant "estimated structural slope parameter in the NKPC" for the range of years 1960-2019.

Methodology

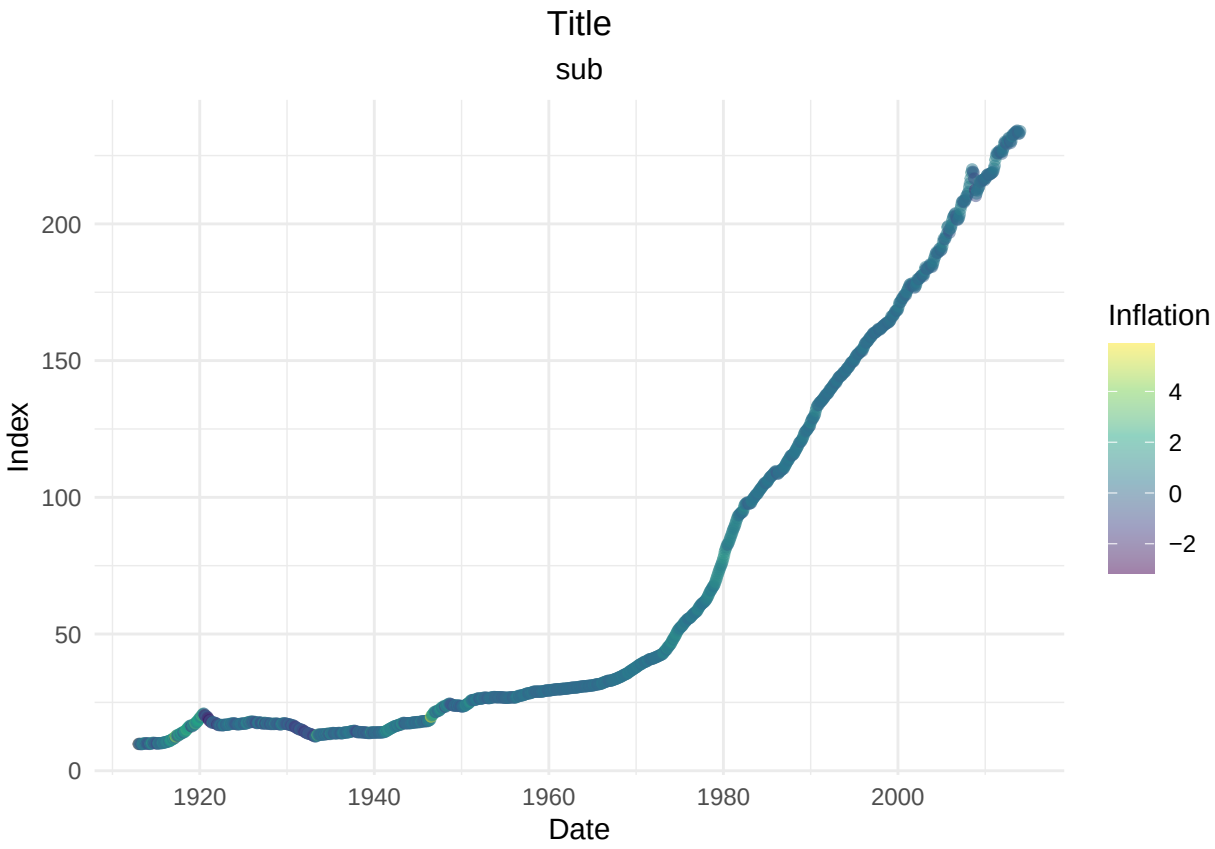
In accordance with studies based on NKPC models, we intend to develop an univariate simple linear regression model for the CPI. This model will differ from others in its segmentation. As previously stated, two stabilized periods of years seem to extend from about 1913 to 1960 and 1980 onwards. We intend to divulge the relationship of these two segmented periods as their own separate models.

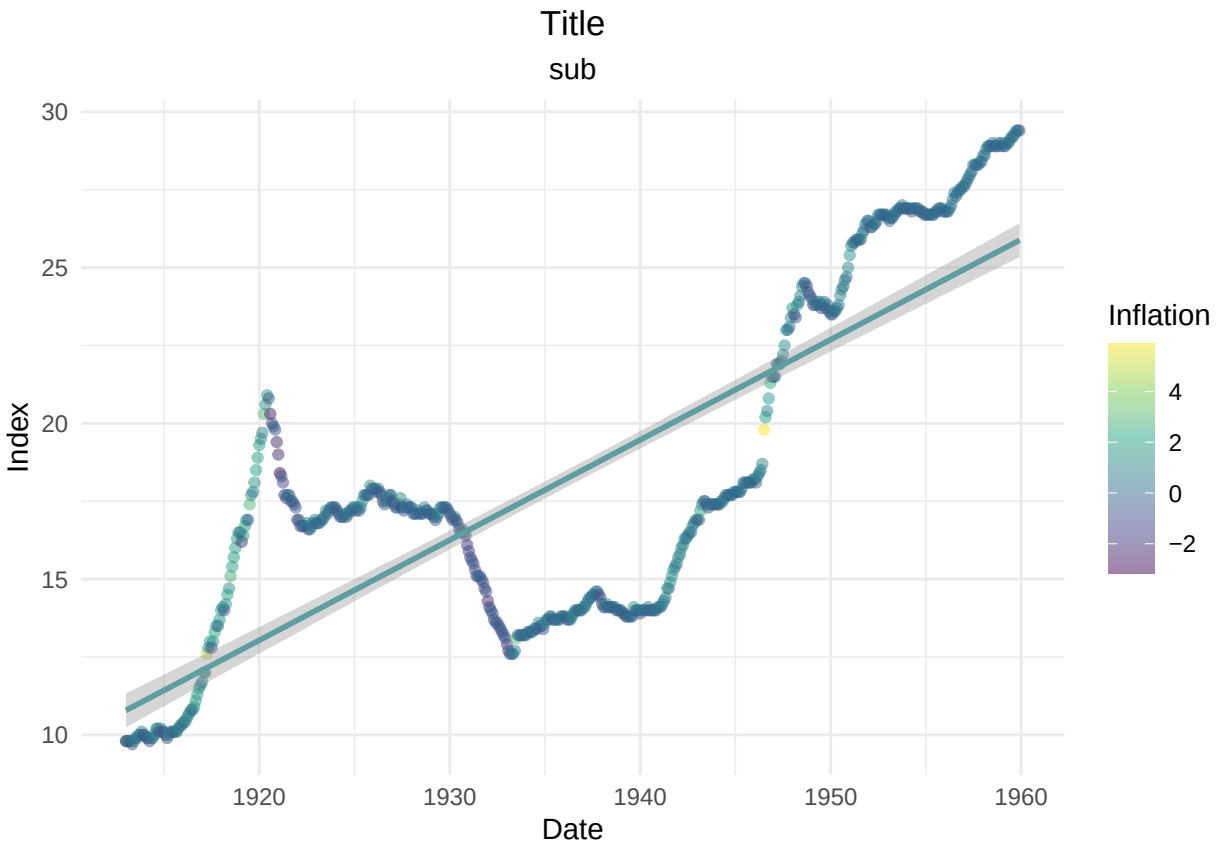
Data is sourced from the BLS and contains monthly CPI estimates for each year over the range 1913-2014. Data collection did not begin until 1913 via authorization from Congress. Monthly estimates are given on the first of each month and start with January 1, 1913. The final monthly estimate ends on January 1, 2014. Inflation is calculated alongside CPI as a new variable but remember, this is only a representative estimate of the change in CPI values from one month to the next.

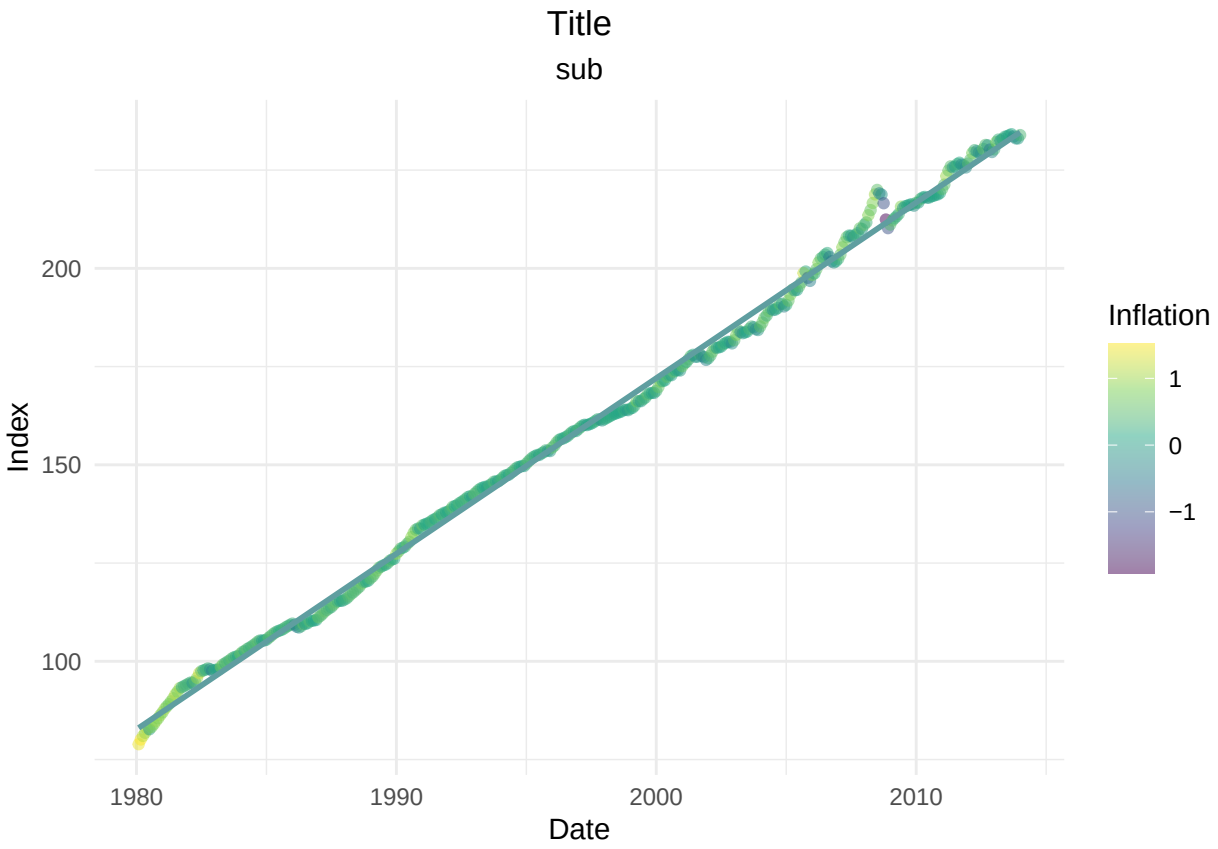
Experimentation and Results

```
## # A tibble: 6 x 3
##   Date       Index Inflation
##   <date>     <dbl>     <dbl>
## 1 1913-01-01   9.8     NA
## 2 1913-02-01   9.8      0
## 3 1913-03-01   9.8      0
## 4 1913-04-01   9.8      0
## 5 1913-05-01   9.7     -1.02
## 6 1913-06-01   9.8      1.03
```

```
cpi %>%
  ggplot(aes(Date, Index)) + geom_point(aes(color=Inflation)) +
  scale_color_viridis_c(aesthetics = c("colour", "fill"), option = "D", alpha = .5) +
  labs(title = "Title", subtitle = "sub", xlab="xlab", ylab="ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5))
```


Title
sub
Index
200
150
100
50
0
1920
1940
1960
1980
2000
Date
Inflation
4
2
0
−2

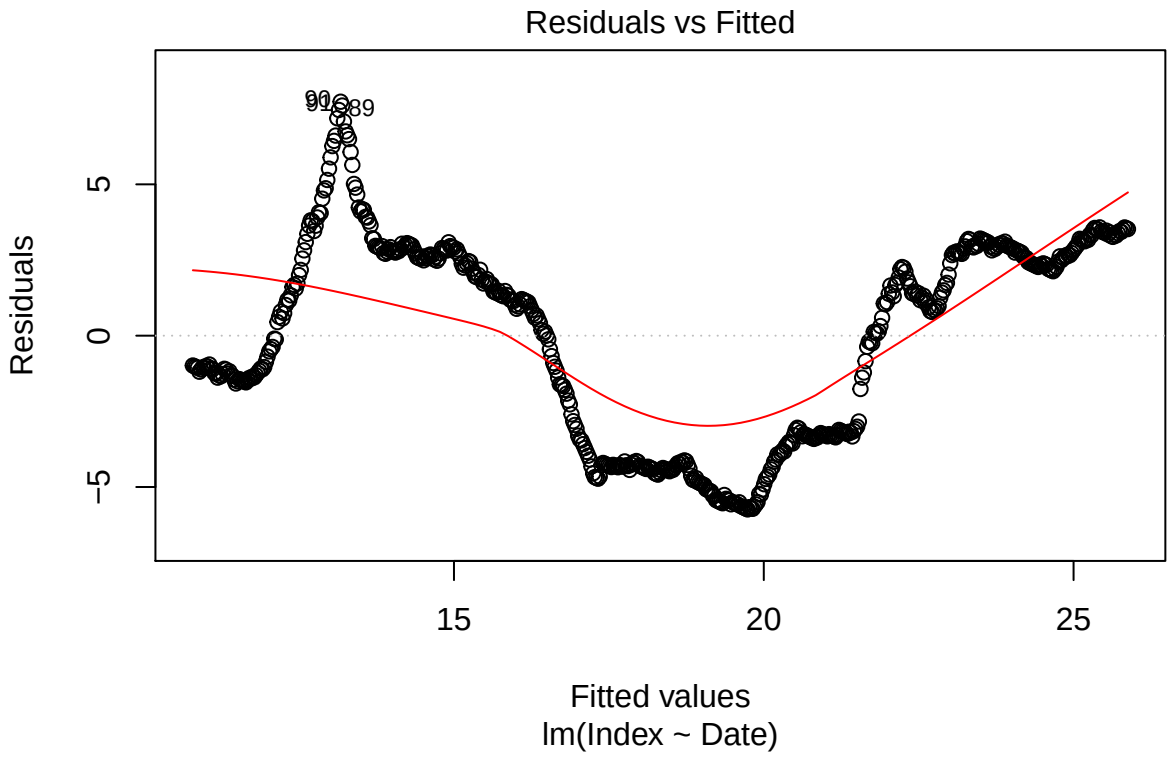

Title
sub
Index
Date
30
25
20
15
10
1920
1930
1940
1950
1960
Inflation
4
2
0
−2

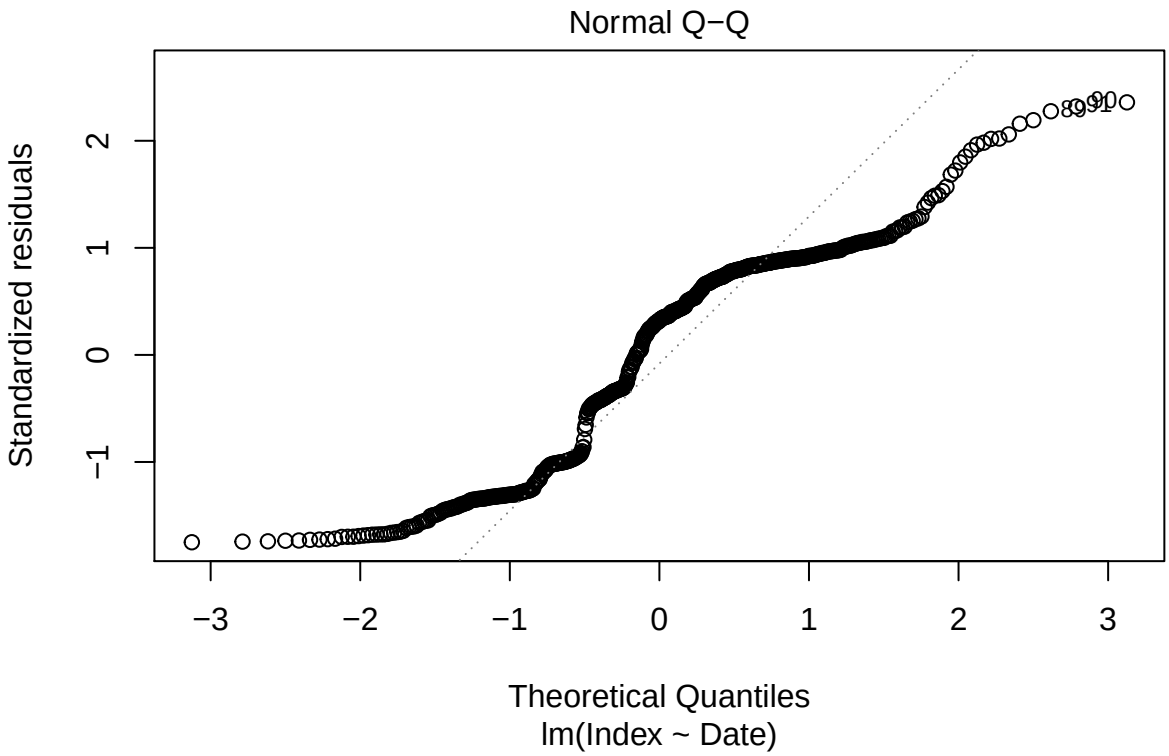


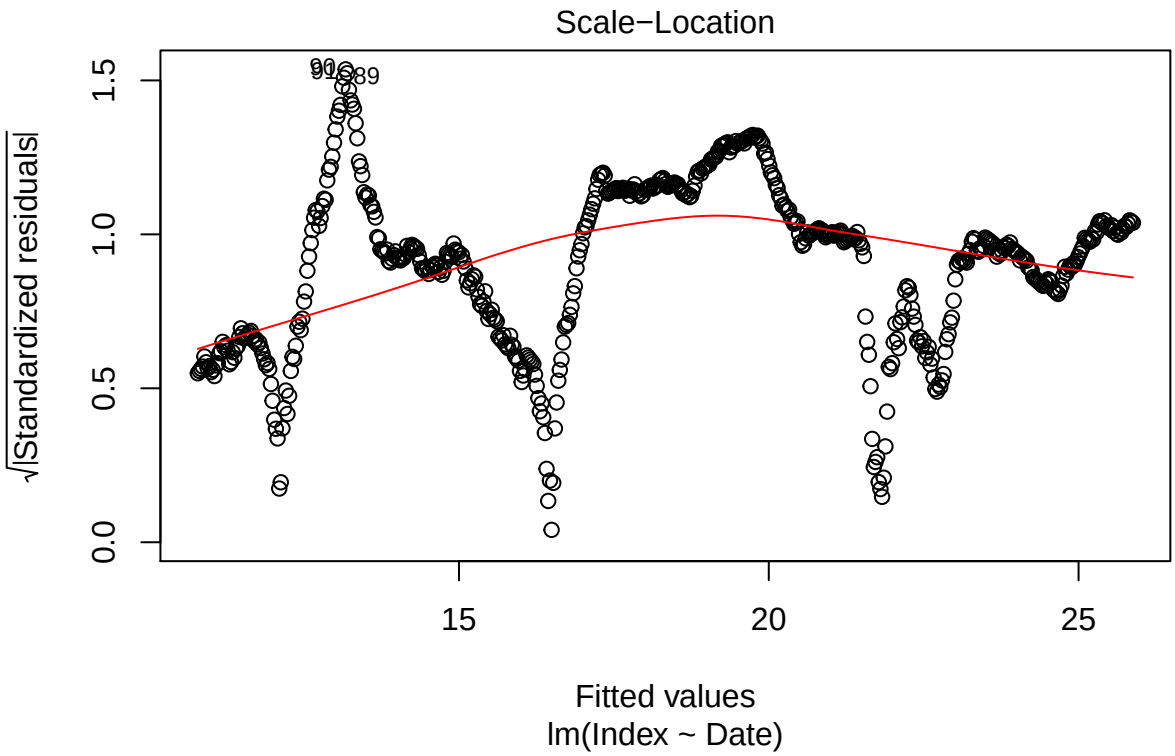

```
g1 <- cpi %>%
  filter(Date < "1960-01-01")
g1.lm <- lm(Index ~ Date, g1)
summary(g1.lm)
```

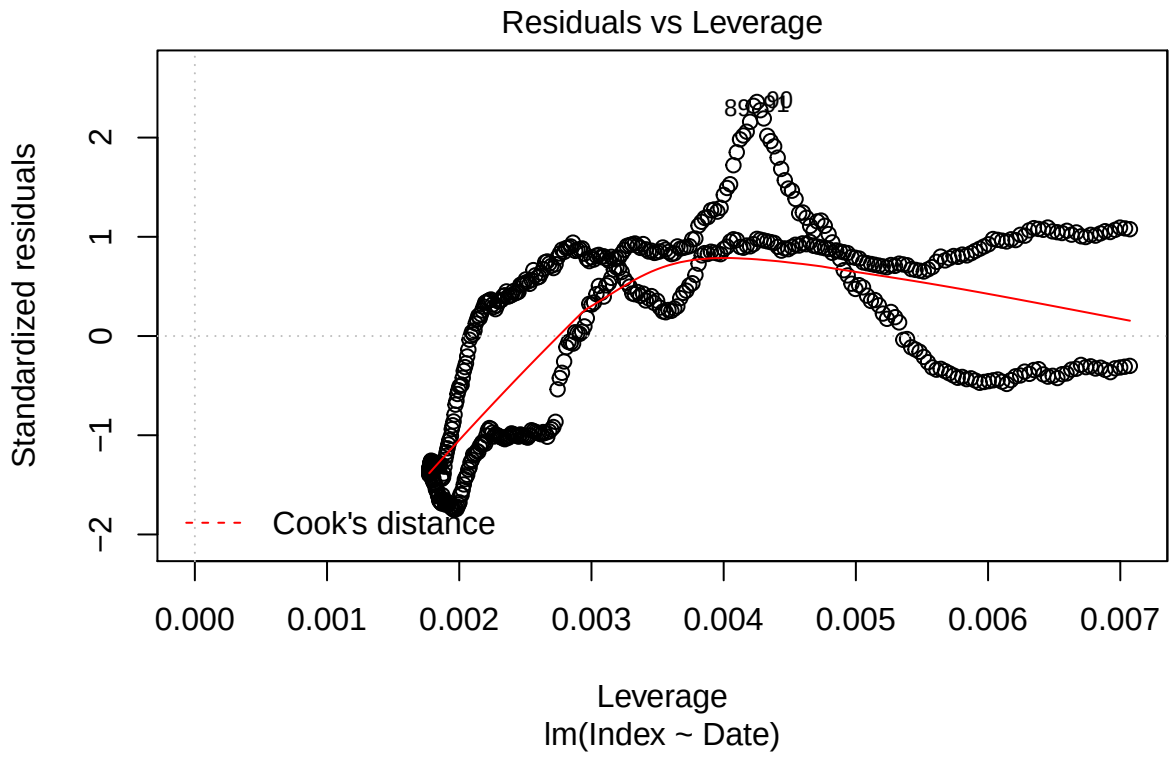
```
##
## Call:
## lm(formula = Index ~ Date, data = g1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.739 -3.305  1.050  2.773  7.729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.912e+01  3.687e-01   78.98   <2e-16 ***
## Date        8.808e-04  2.790e-05   31.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.284 on 562 degrees of freedom
## Multiple R-squared:  0.6394, Adjusted R-squared:  0.6388
## F-statistic: 996.6 on 1 and 562 DF,  p-value: < 2.2e-16
```

```
plot(g1.lm)
```


Residuals vs Fitted
Residuals
5
0
−5
15
20
25
Fitted values
lm(Index ~ Date)


Normal Q−Q
Standardized residuals
−1
0
1
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Index ~ Date)

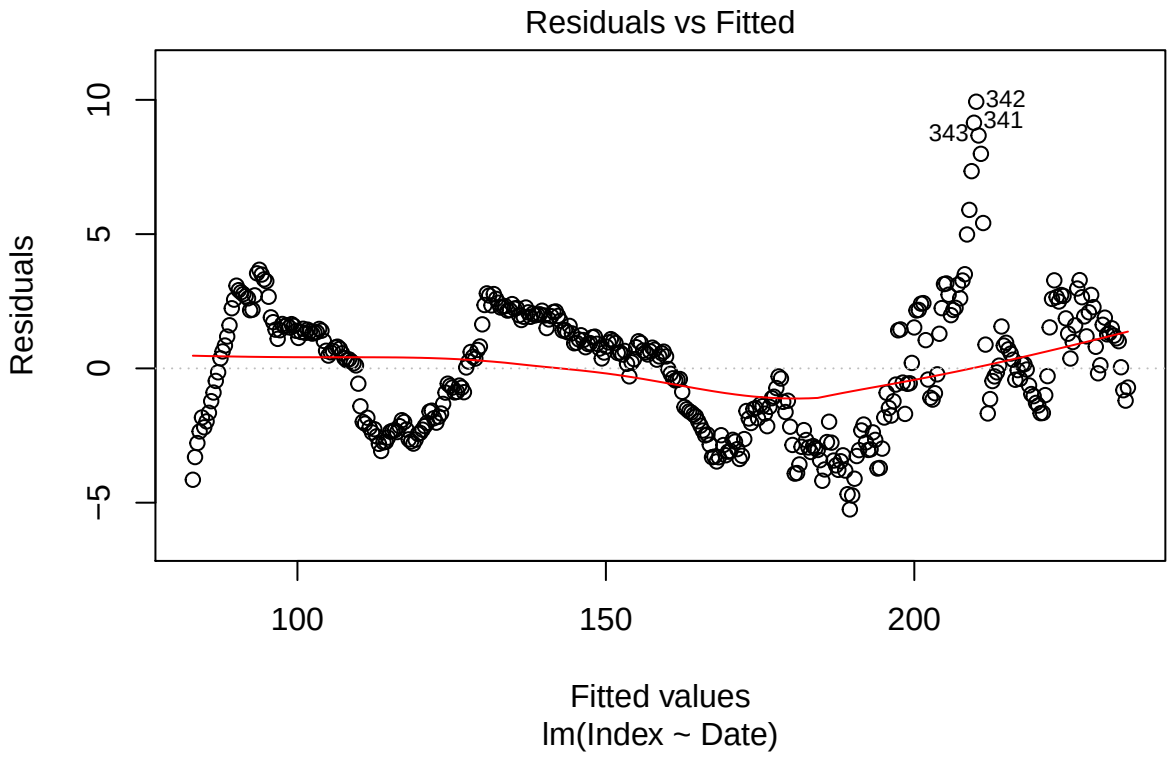

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
Fitted values
lm(Index ~ Date)

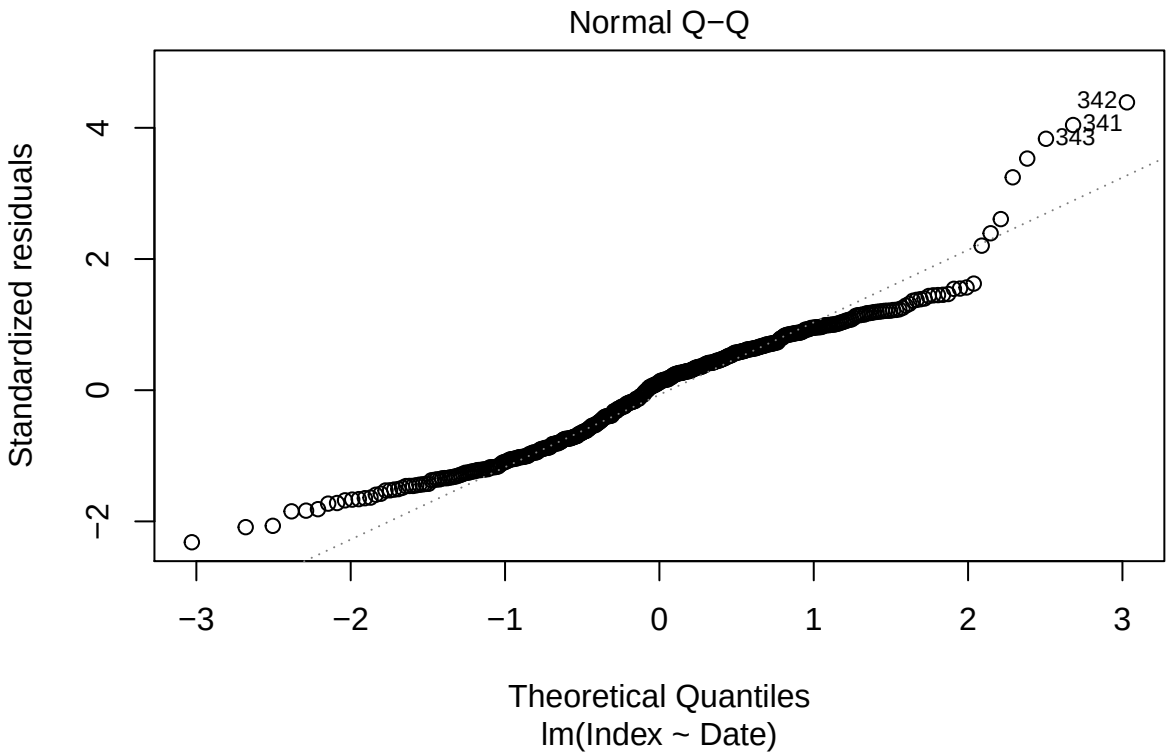


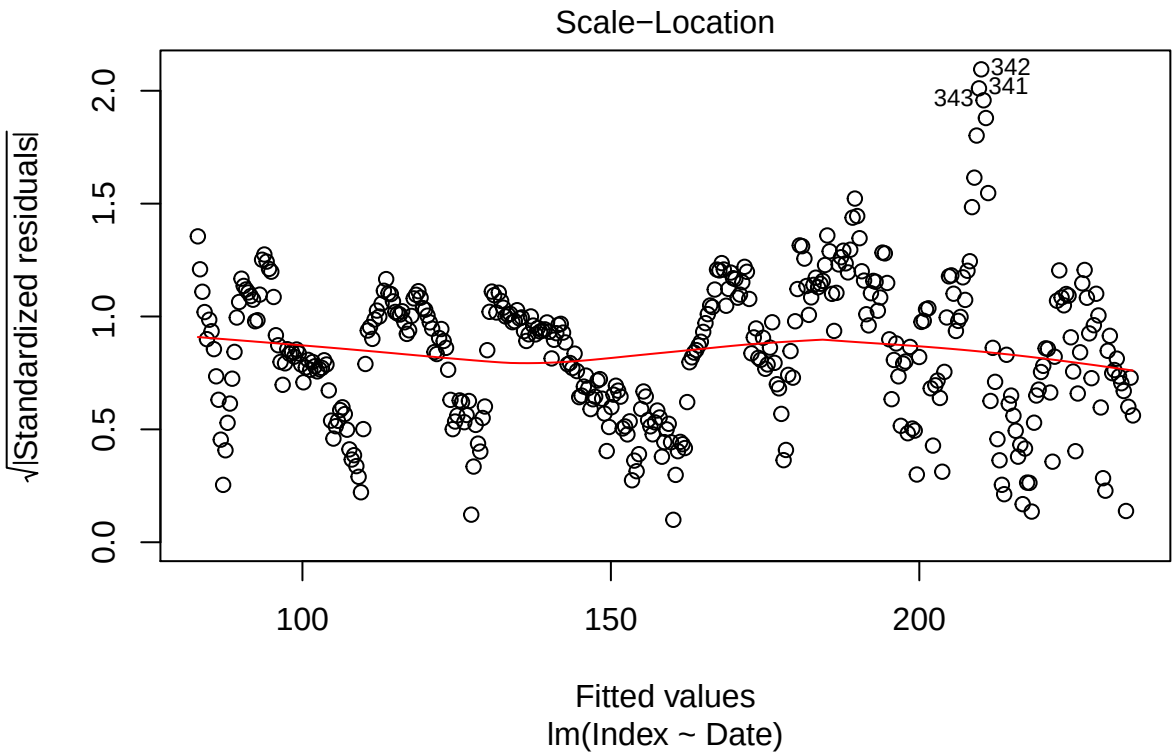

```
g2 <- cpi %>%
  filter(Date > "1980-01-01")
g2.lm <- lm(Index ~ Date, g2)
summary(g2.lm)
```

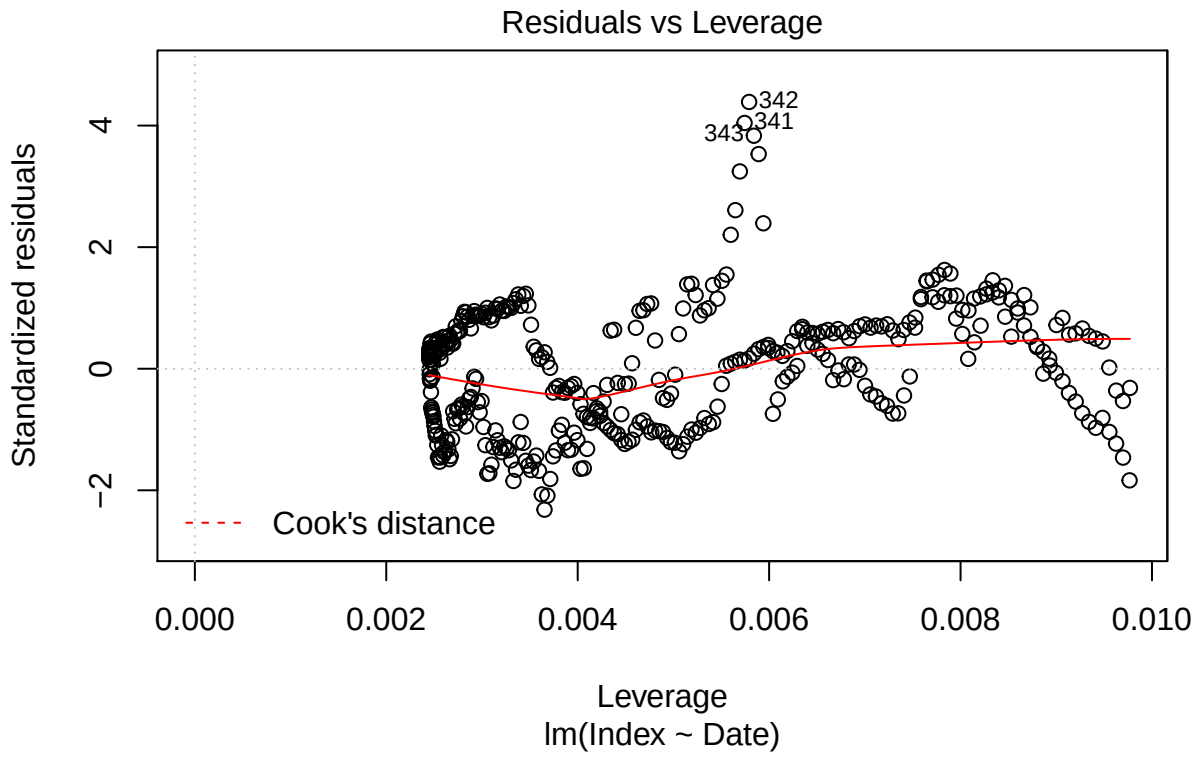
```
##
## Call:
## lm(formula = Index ~ Date, data = g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2504 -1.8410  0.2805  1.5275  9.9303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.798e+01  3.293e-01   115.3   <2e-16 ***
## Date        1.224e-02  3.134e-05   390.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 406 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9973
## F-statistic: 1.524e+05 on 1 and 406 DF,  p-value: < 2.2e-16
```

```
plot(g2.lm)
```


Residuals vs Fitted
Residuals
10
5
0
−5
342
341
343
100
150
200
Fitted values
lm(Index ~ Date)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Index ~ Date)
342
341
343
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
342
341
343
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
150
200
Fitted values
lm(Index ~ Date)



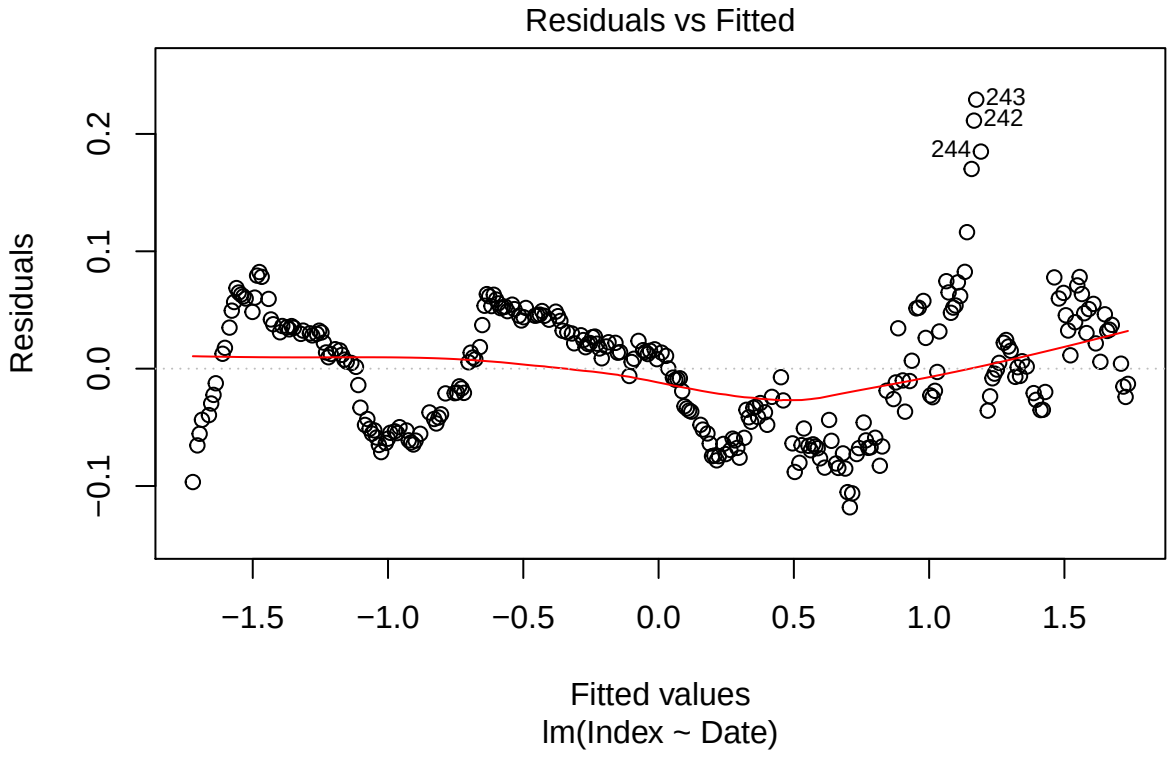

```
# Split 70-30
set.seed(41)
tindex <- createDataPartition(g2$Index, p = .7, list = FALSE, times = 1)
train <- g2[tindex,]
test <- g2[-tindex,]
# Determine best transformation
bestNormalize(train$Index)
```

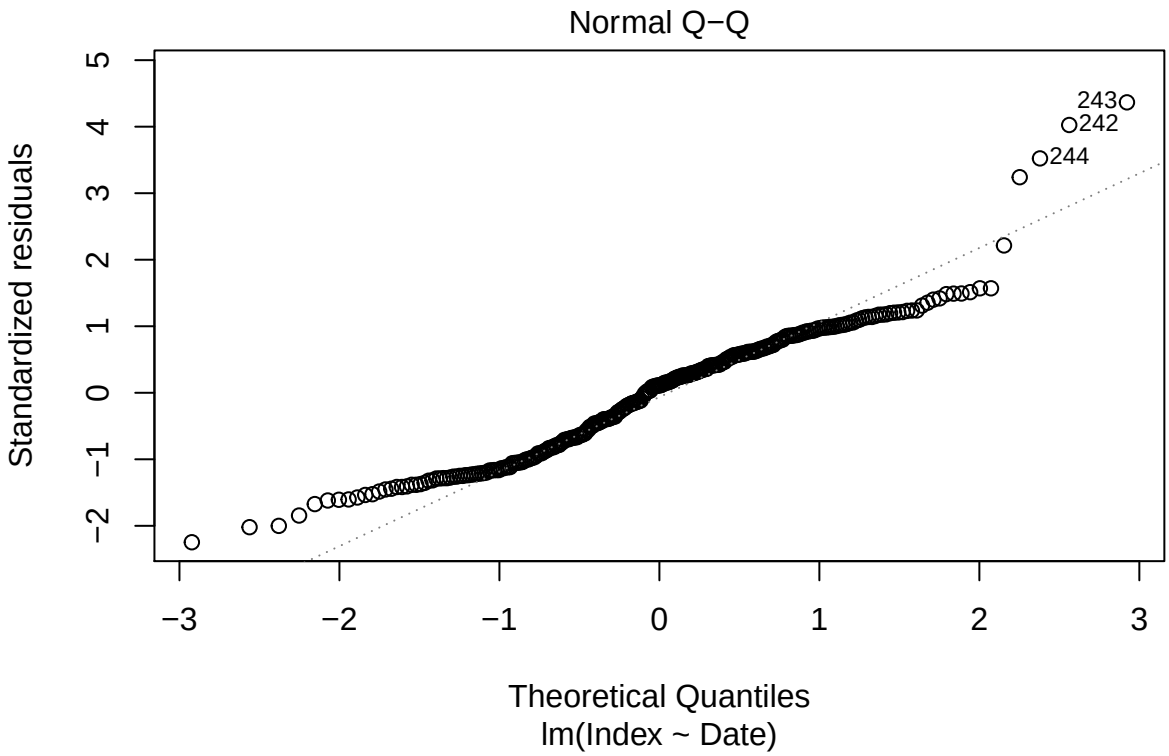
```
## Best Normalizing transformation with 288 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - arcsinh(x): 1.3294
##  - Box-Cox: 1.2658
##  - Center+scale: 1.1924
##  - Exp(x): 34.7778
##  - Log_b(x+a): 1.3294
##  - orderNorm (ORQ): 1.2422
##  - sqrt(x + a): 1.2394
##  - Yeo-Johnson: 1.268
## Estimation method: Out-of-sample via CV with 10 folds and 5 repeats
##
## Based off these, bestNormalize chose:
## center_scale(x) Transformation with 288 nonmissing obs.
##  Estimated statistics:
##  - mean (before standardization) = 158.5135
##  - sd (before standardization) = 43.79097
```

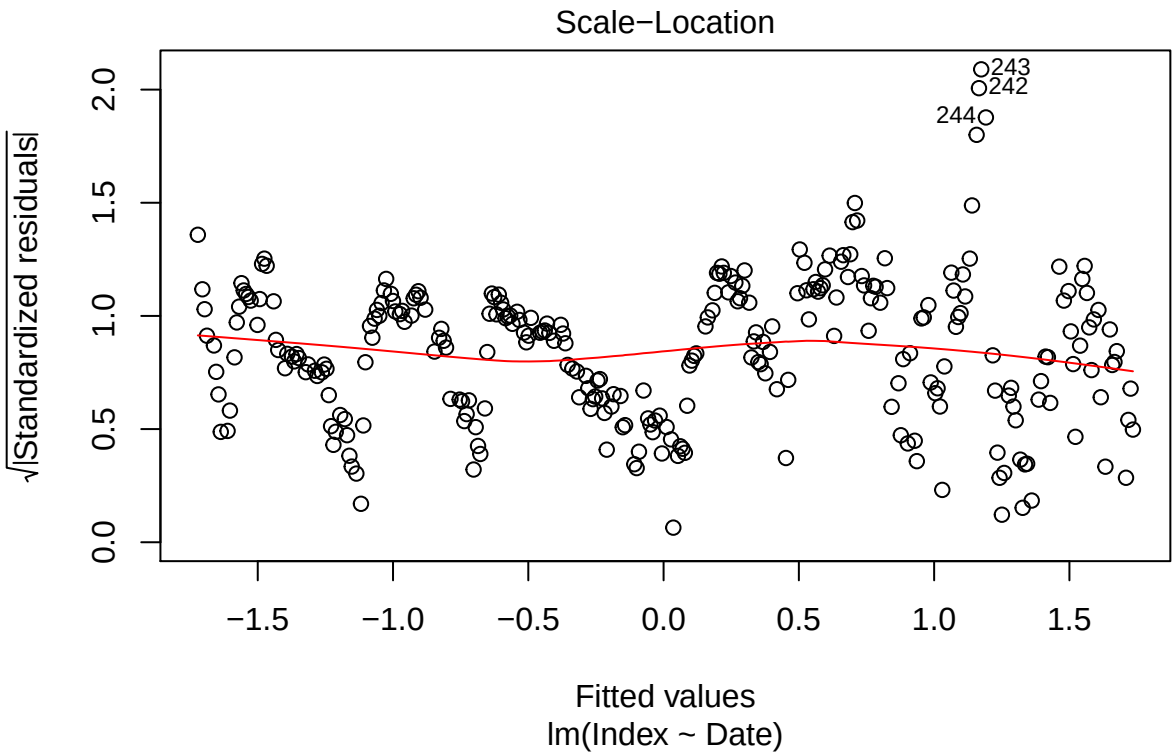
```
train$Index <- scale(train$Index)
scaled.mod <- lm(Index ~ Date,train)
summary(scaled.mod)
```

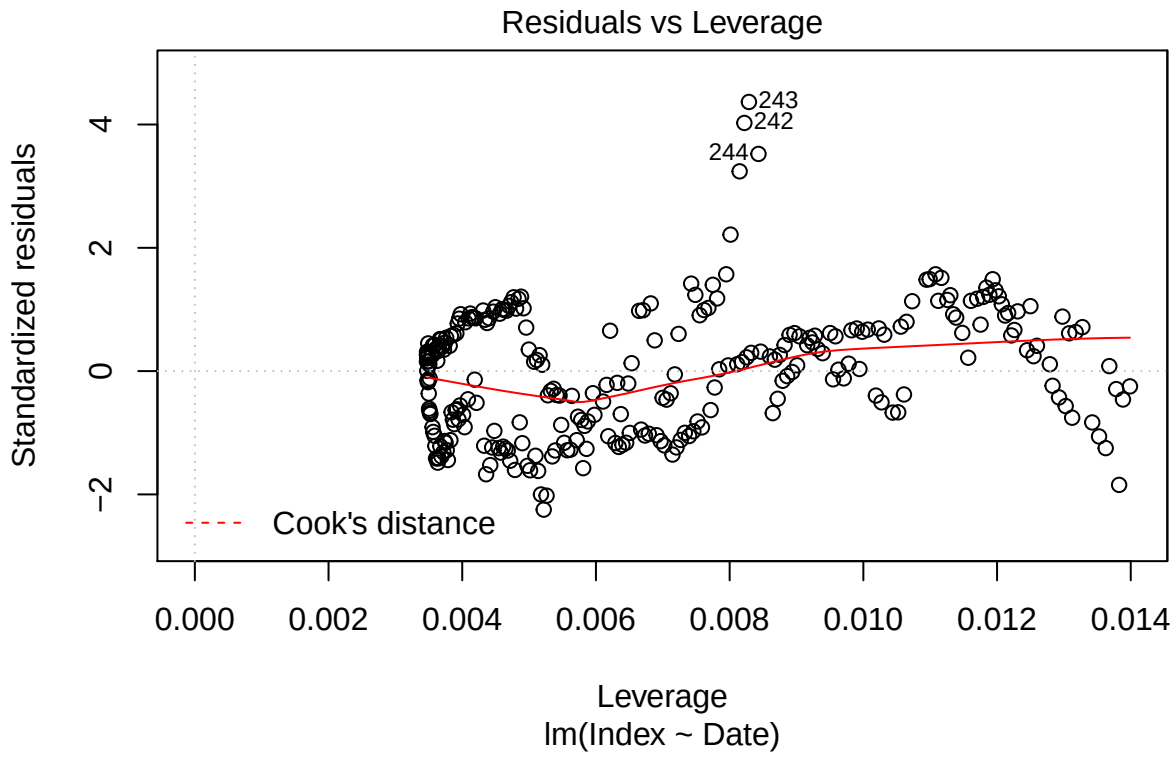
```
##
## Call:
## lm(formula = Index ~ Date, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118151 -0.043081  0.005851  0.036281  0.229216
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.749e+00  9.113e-03  -301.7   <2e-16 ***
## Date         2.790e-04  8.695e-07   320.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05272 on 286 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9972
## F-statistic: 1.03e+05 on 1 and 286 DF,  p-value: < 2.2e-16
```

```
plot(scaled.mod)
```

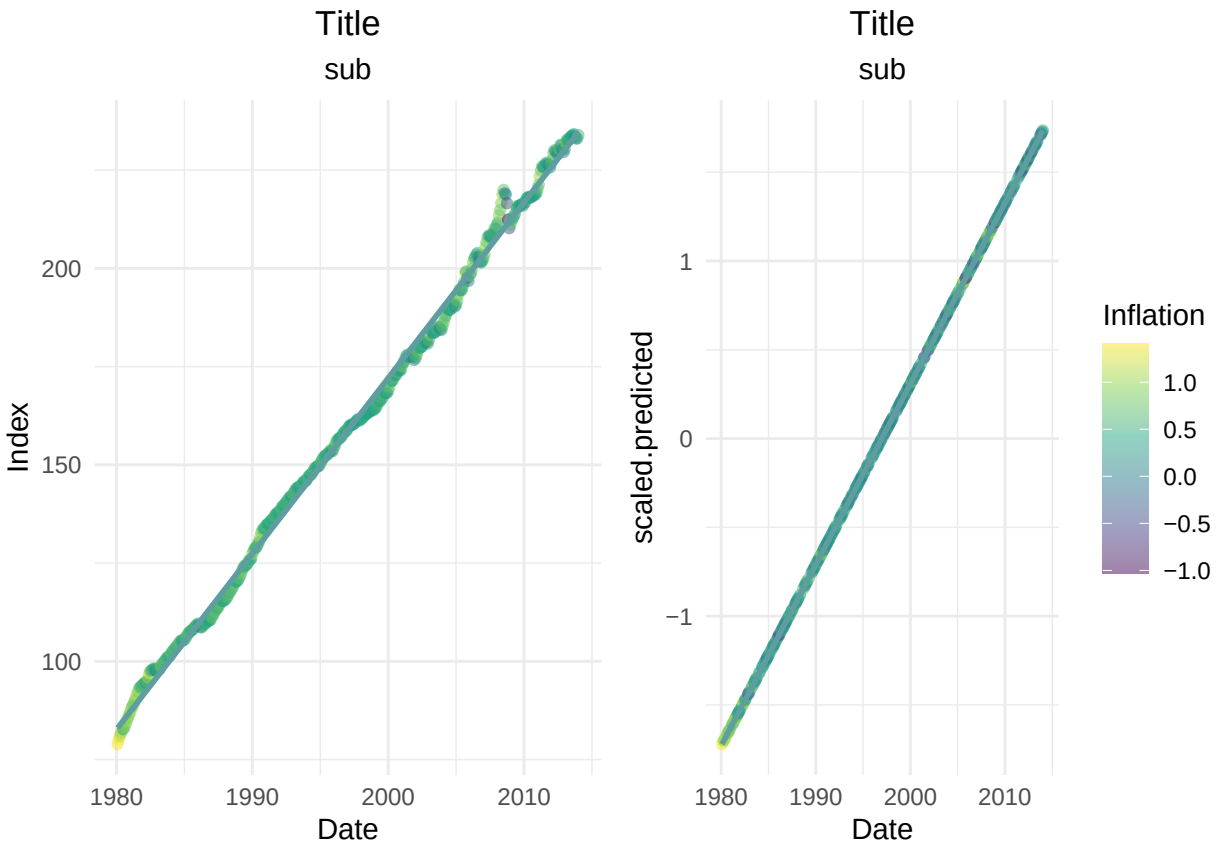



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Index ~ Date)
243
242
244
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
3
4
5


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
243
242
244
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Fitted values
lm(Index ~ Date)




```
# Calculate prediction values
scaled.predicted <- predict(scaled.mod, train)
# Add them to the data frame
train$scaled.predicted <- scaled.predicted
# Calculate errors, magnitude, rmse
train <- train %>%
  mutate(scaled.error = Index - scaled.predicted) %>%
  mutate(scaled.mag = scaled.error^2) %>%
  mutate(scaled.eravg = mean(scaled.mag)) %>%
  mutate(scaled.rmse = sqrt(scaled.eravg))
```




```
# Visualize error type
scaled.mag.dense <- train %>%
  ggplot(aes(scaled.mag)) +
  geom_density(aes()) +
  labs(subtitle = "sub") +
  xlab("xlab") +
  ylab("ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5), legend.positio
scaled.mag.hist <- train %>%
  ggplot(aes(scaled.mag)) +
  geom_histogram(aes()) +
  labs(subtitle = "sub") +
  xlab("xlab") +
  ylab("ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5), legend.positio
scaled.error.hist <- train %>%
  ggplot(aes(scaled.error)) +
  geom_histogram(aes()) +
  labs(subtitle = "sub") +
  xlab("xlab") +
  ylab("ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5), legend.positio
scaled.error.dense <- train %>%
  ggplot(aes(scaled.error)) +
  geom_density(aes()) +
  labs(subtitle = "sub") +
```

```
  xlab("xlab") +
  ylab("ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5), legend.positio
ggarrange(scaled.error.dense, scaled.mag.dense, scaled.error.hist, scaled.mag.hist)
```