CPI Analysis

Zachary Palmore

5/16/2021

Abstract

Businesses, governments, and consumers in the United States all rely on some aspect of the consumer price index (CPI) when making financial decisions. There is evidence to suggest that improperly measuring this statistic by margins as small as 1.1% could result in trillions of dollars lost through government spending. This notion ties into the economic behavior of businesses and consumers quite intuitively and yet, we rely on models that predict using data on business and consumer behavior that largely differs in our society today. Furthermore, recent studies have identified a stabilization in the rate of inflation (which is found through the CPI) since about 1980. This phenomenon has created a dichotomy in the rate of change in prices of consumer goods and services in which CPI estimates earlier than 1960 follow one pattern and the estimates after 1980 follow another separate pattern. Through univariate simple linear regression techniques we developed a contemporary model that exploits the dichotomous nature of these statistics and we attempt to prove that the holistic use of historical data does not accurately capture the behavior of modern society nor does it aid in its prediction.

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 has 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 offer 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 in today's economic environment 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 their 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 has 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 a 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. The benchmark for which the entire data set is evaluated is the arithmetic mean of the annual CPI for the 1982-1984. 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. Inflation calculations did not begin until May 1, 2014.

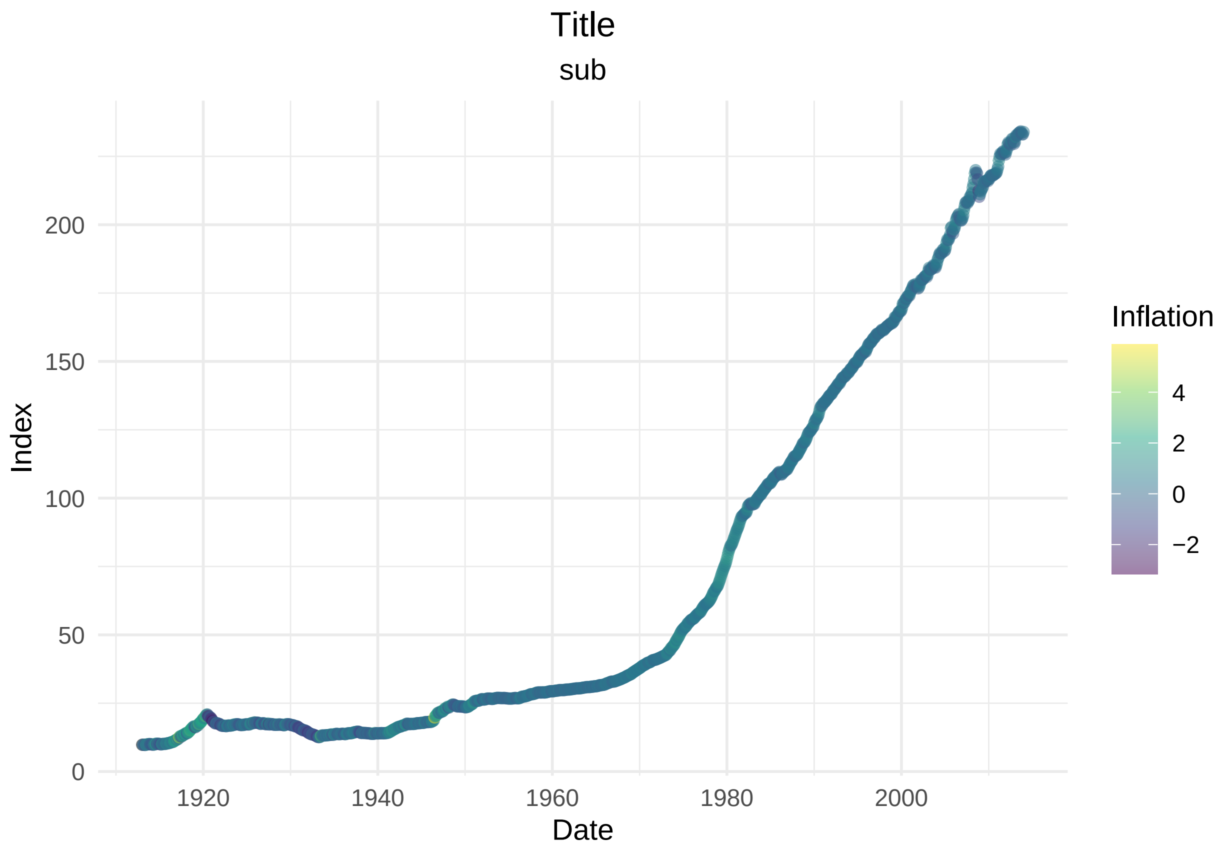
Experimentation and Results

Newer portions of the data were stored in reports released by the BLS corresponding to the timespan between the first of May in 2014 to the first of March in 2021. This data would only be used in evaluation of the model performance statistics. All preceding data from the first of May in 2014 backwards, was made available in a comma delimited format and was extracted from the open source API by data.io for hosting purposes and reproducibility. See the appendix for further descriptions of exactly how this data was extracted. The first few observations are shown.

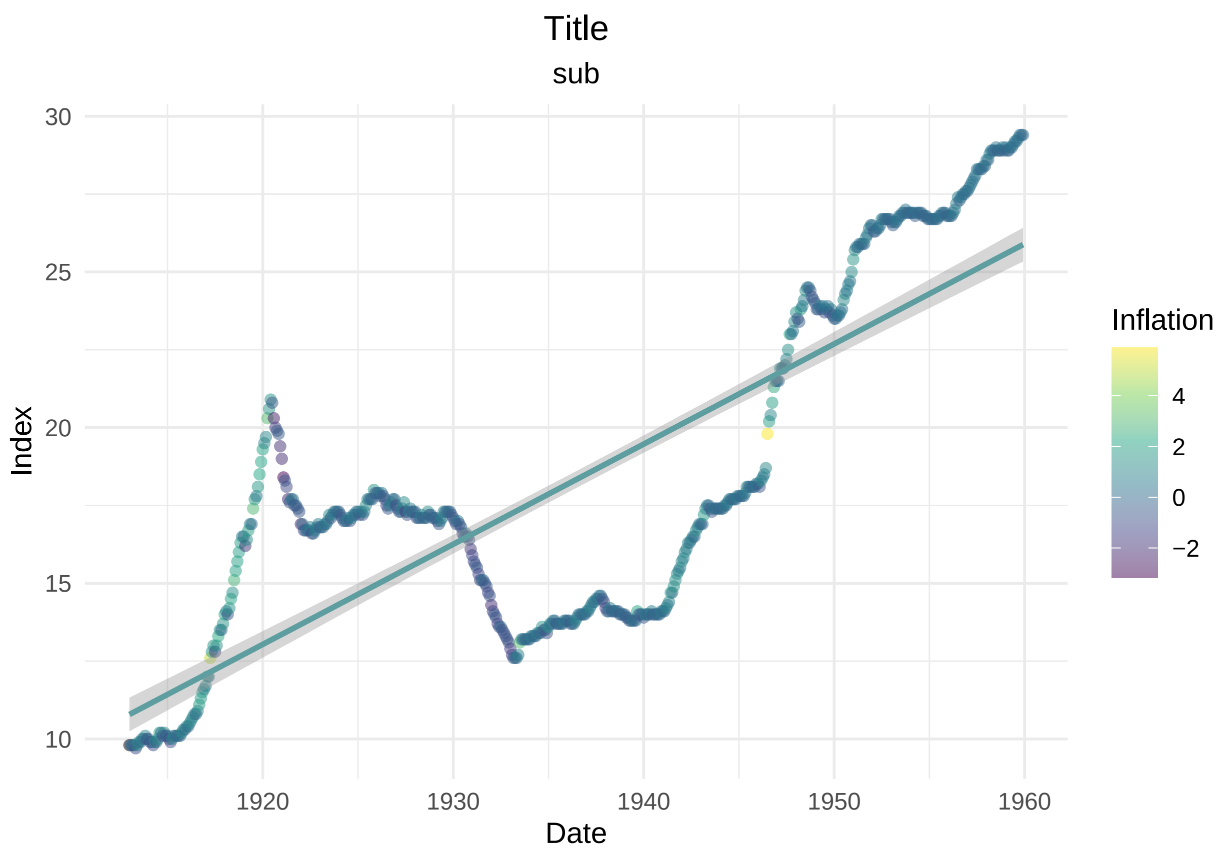
```
## # A tibble: 6 x 3
##   Date       Index Inflation
##   <date>     <dbl>     <dbl>
## 1 1913-01-01   9.8     NA
## 2 1913-02-01   9.8      0
## 3 1913-03-01   9.8      0
## 4 1913-04-01   9.8      0
## 5 1913-05-01   9.7     -1.02
## 6 1913-06-01   9.8      1.03
```

Our first step in this experiment is the separation of the data into two distinct segments, and to do so we first justify the need and approximate the periods with which to draw from. This is merely a check on the literature but we can see the results as a scatterplot. To assist in the attempted segmentation based on mathematical behavior, we do not add a trend to the plot at this time. Inflation is highlighted on a continuous scale for each month within the range.

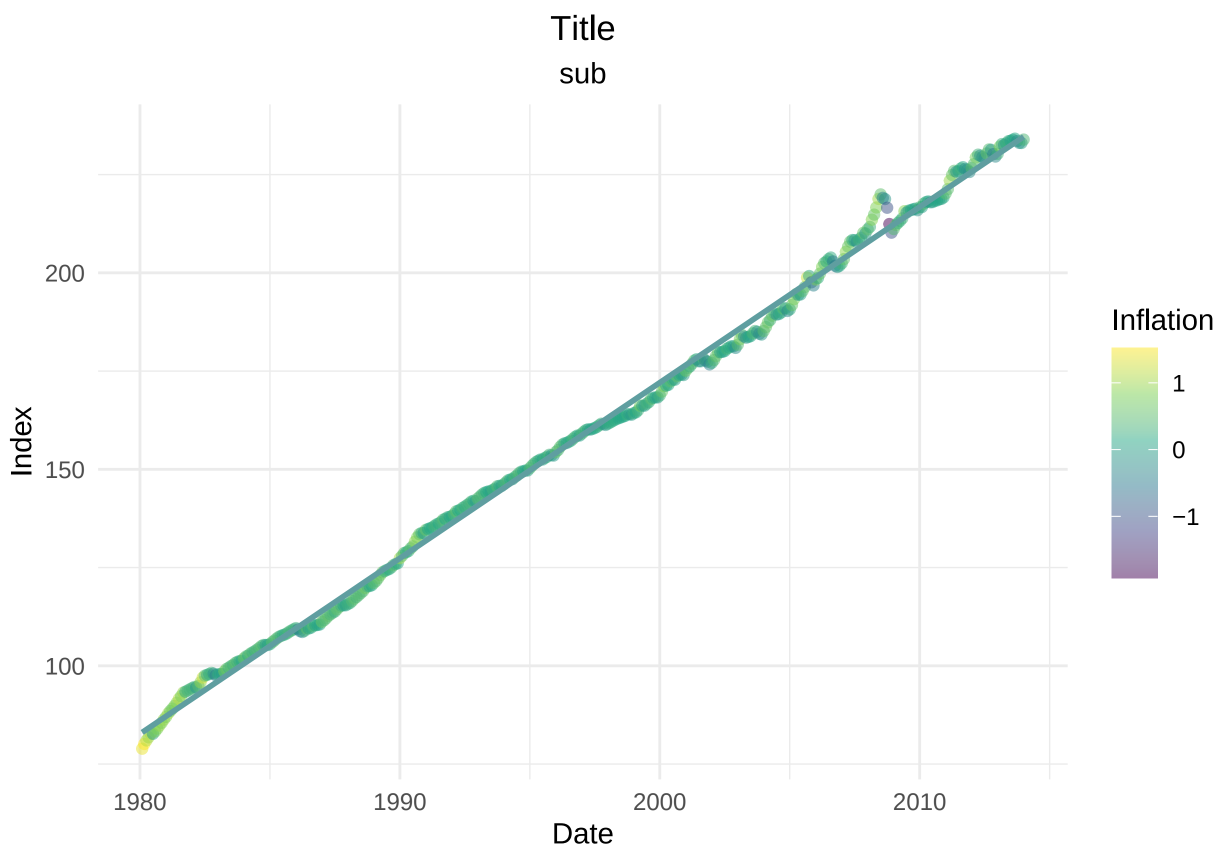
```
cpi %>%
  ggplot(aes(Date, Index)) + geom_point(aes(color=Inflation)) +
  scale_color_viridis_c(aesthetics = c("colour", "fill"), option = "D", alpha = .5) +
  labs(title = "Title", subtitle = "sub", xlab="xlab", ylab="ylab") +
  theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust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



Showing the distribution of CPI from 1913 to 2014 gives us insight into the two groups, because prior to the late 1960's the data appear to fall in line with steady slope of some caliber while the data after the late 1970's follows a much steeper incline. This is our basis for segmentation. Meanwhile, the inflation colored at this scale is practically constant except for a jump in the 1920's and a rapid exponentiation of inflation for the late 1960's to the early 1980's. We examine this further at two closer scales.



By filtering the data to include only those CPI values that were less than the first of January in 1960, a new story emerges. It is clear that the slope is not constant. There is a substantial amount of variation in the CPI and inflation throughout the 1913 - 1960 range. Presumably, this is due to differences in fiscal policy that occurred before 1920 and thereafter when the Great Depression hit the United States. Needless to say, prediction based on a model of this data would result in less than ideal conditions for understanding future CPI and inflation rates.

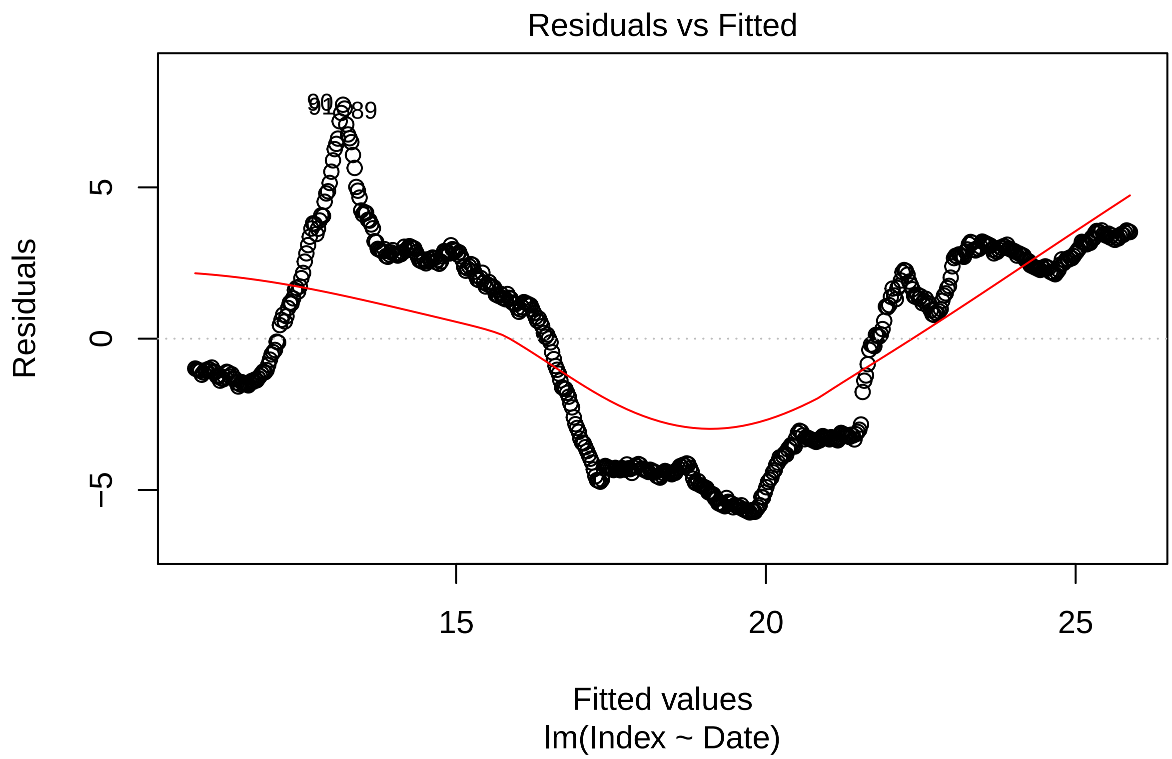


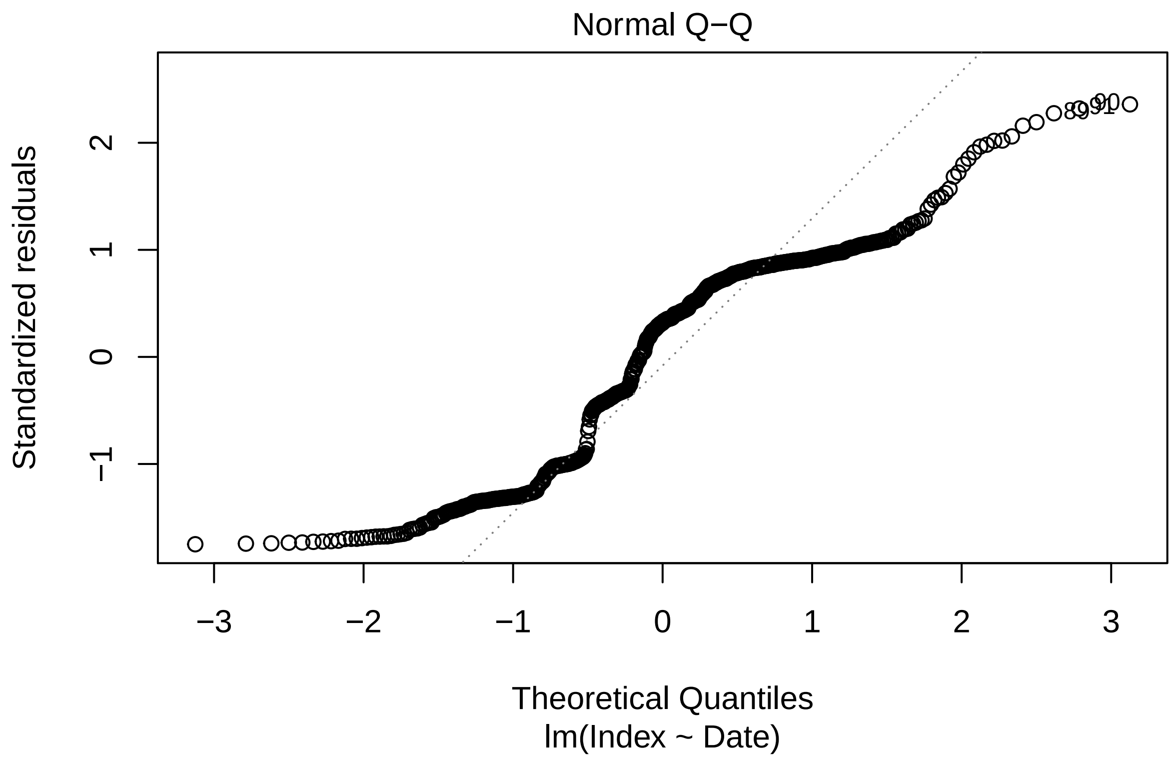
Filtering from the first of January in 1980 until the upper limit of our data range on the first of January in 2014, we get a near constant slope. There is only one minor ‘blip’ or divergence from this expected trend around 2008. Arguably, this is a better segment to make predictions and draw recent conclusions from. A major cause of the resultant blip is likely the Great Recession which certainly resulted in an intense fluctuation of many consumer goods.

Model development is straightforward using the segmentation date ranges as selection, or specifically filtering criteria for our first pre-seventies date range to run through our simple linear regression. We then run diagnostic plots to evaluate our assumptions. Given the somewhat poor linearity we should expect a poorly fitted residual vs fitted plot, if there is a well enough presence at all. Judging from the previous scatterplot and linear regression overlay, we know very few CPI values were solid hits on the trend. A valid normality may also be a concern.

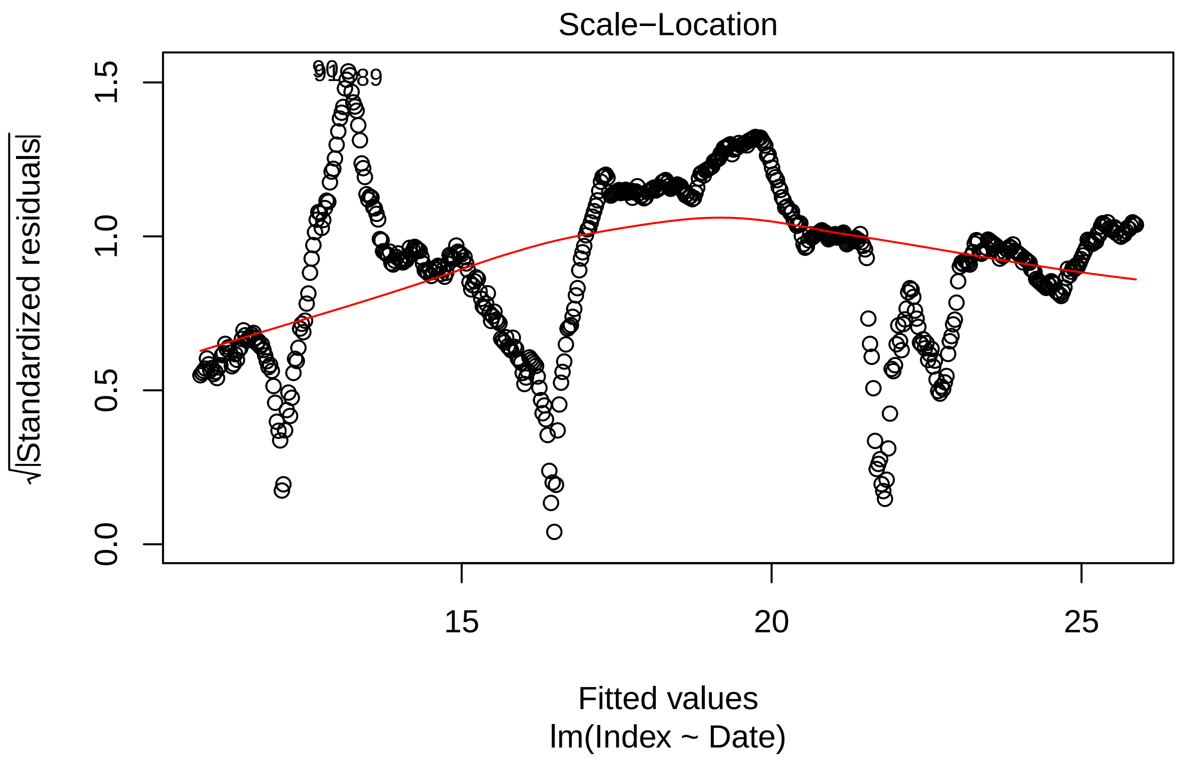
```
##
## Call:
## lm(formula = Index ~ Date, data = g1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.739 -3.305  1.050  2.773  7.729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.912e+01  3.687e-01   78.98   <2e-16 ***
## Date        8.808e-04  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
```

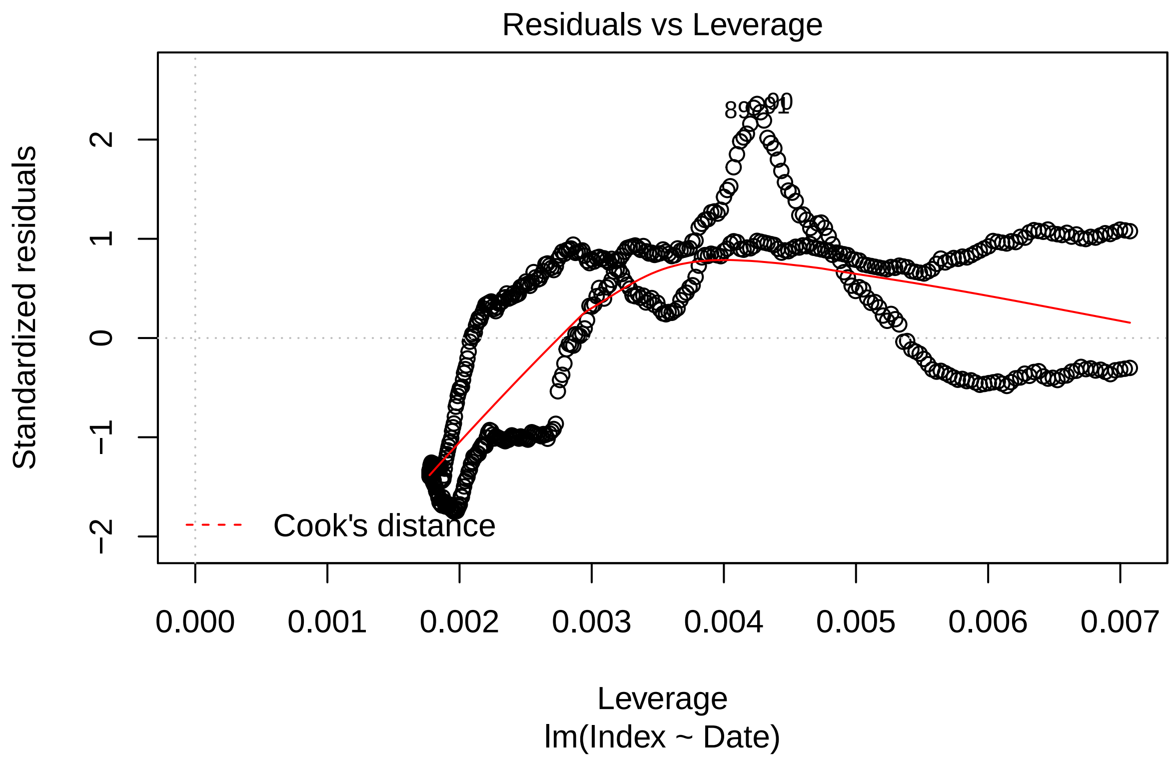




Normal Q−Q
Standardized residuals
−1
0
1
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Index ~ Date)



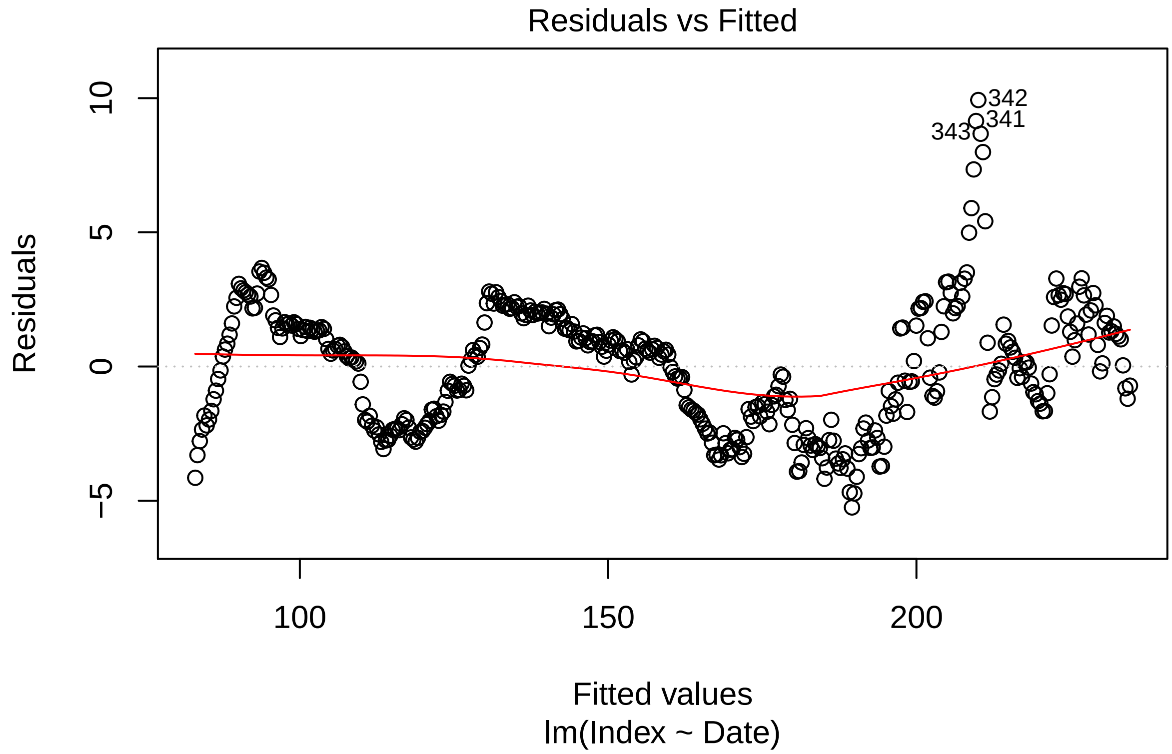
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
Fitted values
lm(Index ~ Date)
90
91
89

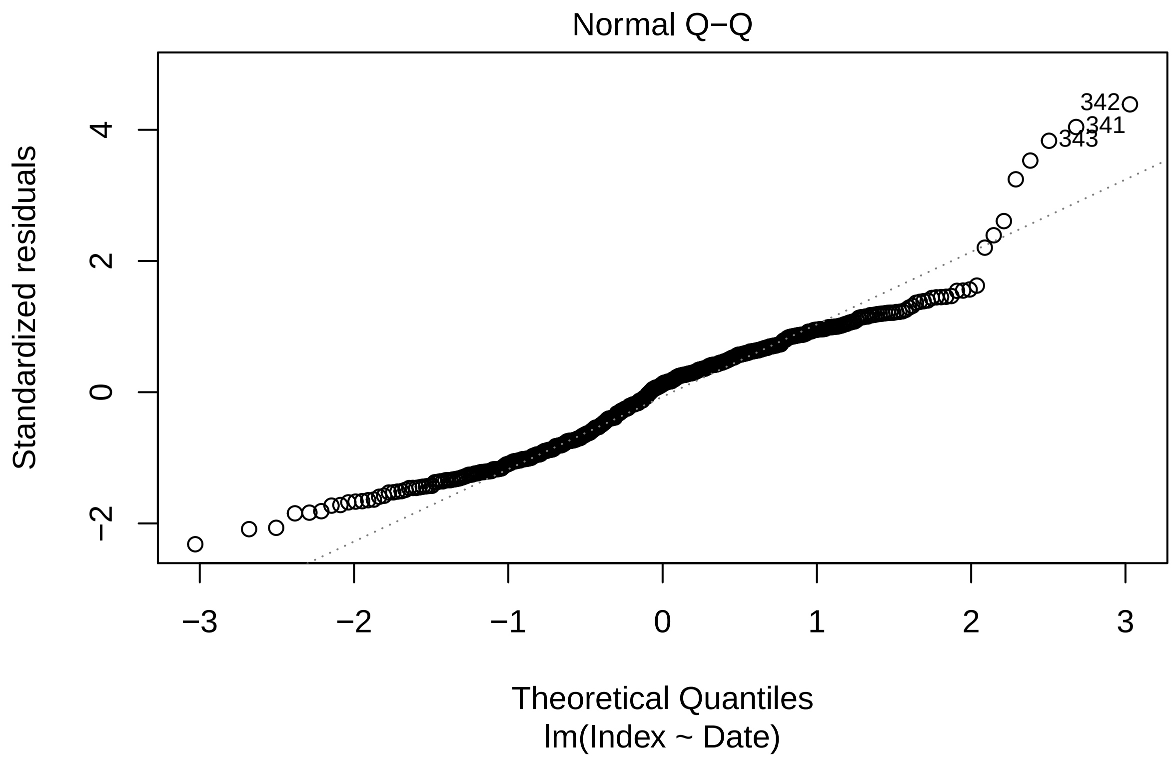


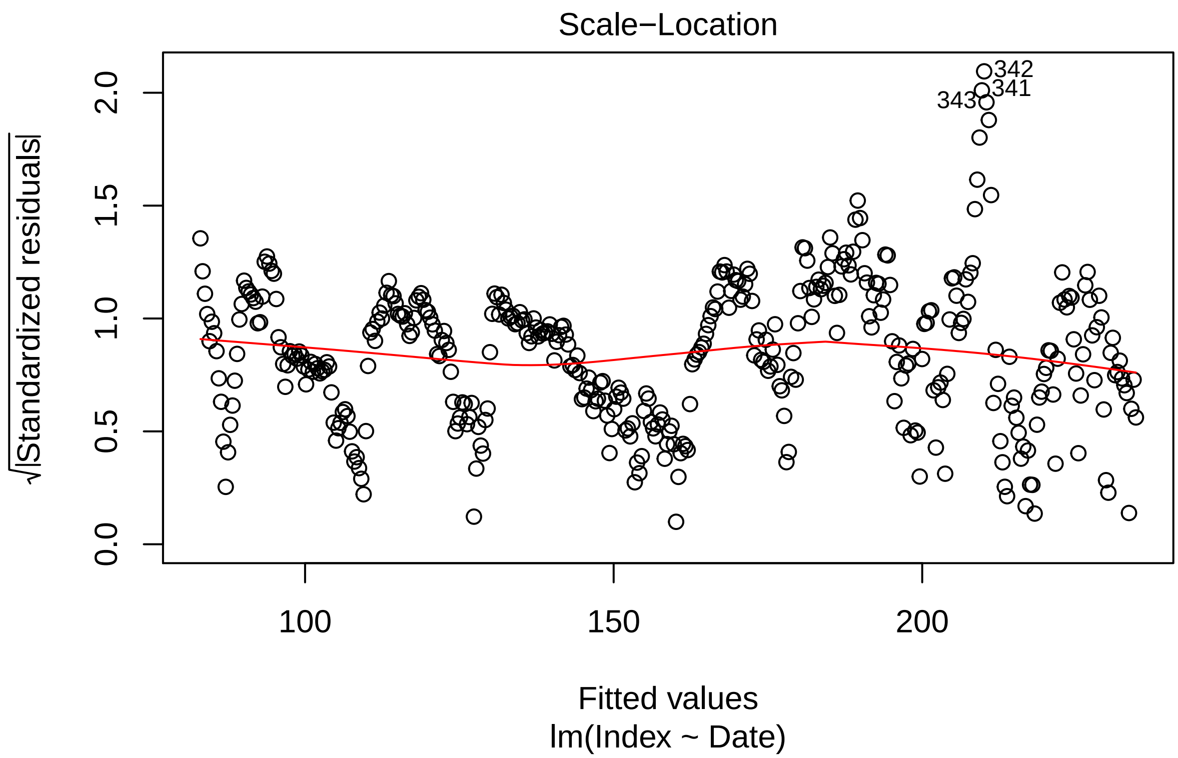
As expected, linearity is poor and normality is not present. While we could attempt to overlook the normality of the plot through various transformations, such efforts would be wasted as there is likely no reasonable transformation that would not completely change the meaning of the data. When even the Scale-Location and Residual vs Leverage plot indicates the presence of highly influential values, there is little hope to piecemeal the variations from this range into something practical. Instead, we should make note of this pattern as it confirms our thoughts that data prior to the first of January in 1960 does indeed behave differently than any expectations of linearity. We repeat this preliminary modeling process on the data from 1980 onward.

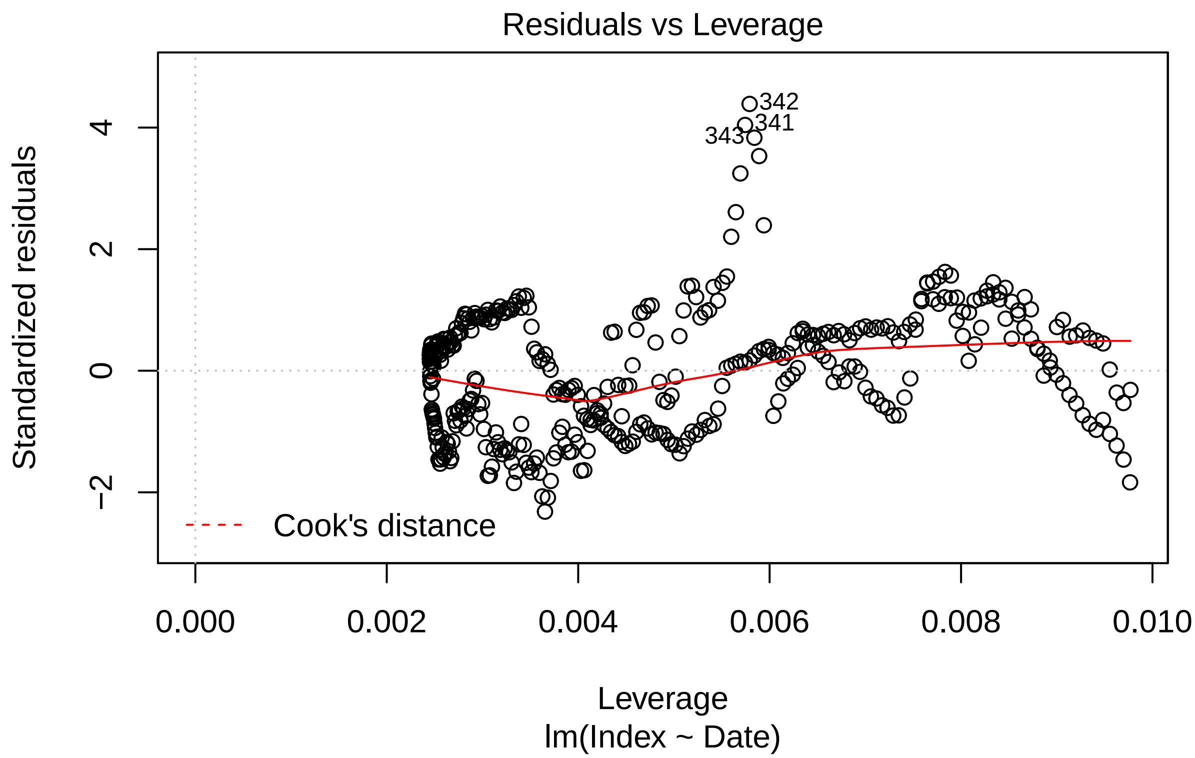
```
##
## Call:
## lm(formula = Index ~ Date, data = g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2504 -1.8410  0.2805  1.5275  9.9303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.798e+01  3.293e-01   115.3   <2e-16 ***
## Date        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
```




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


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

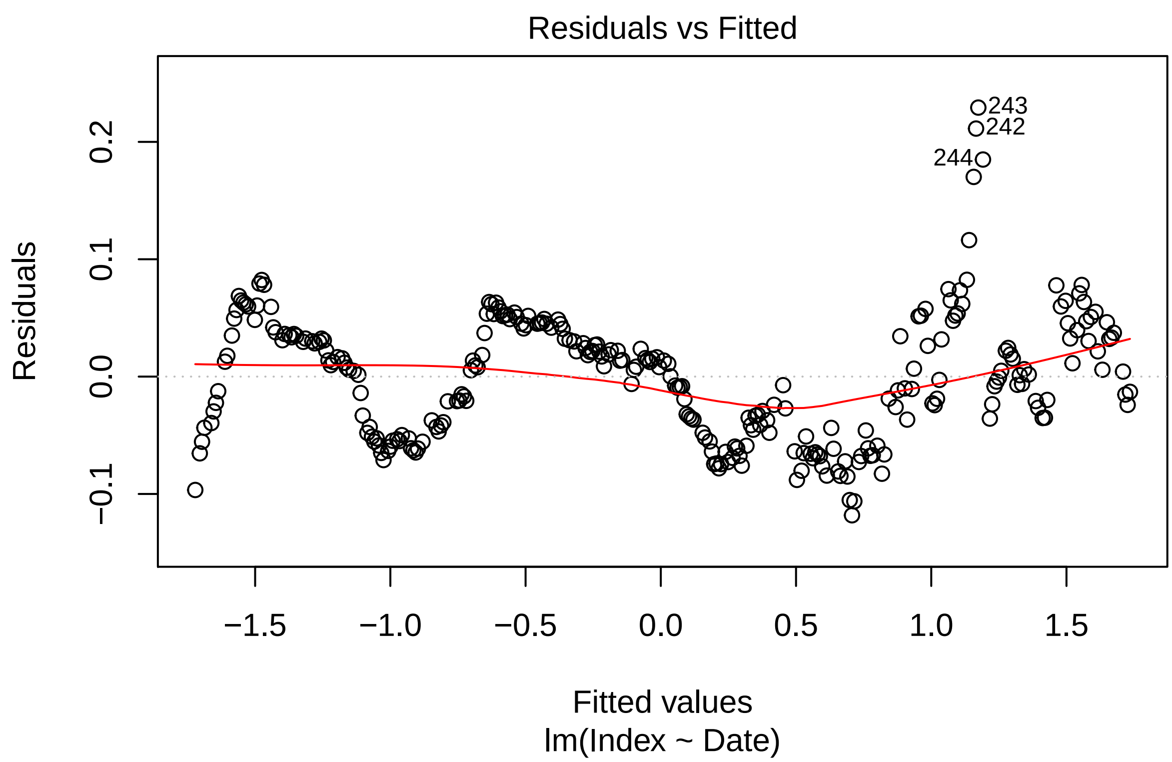


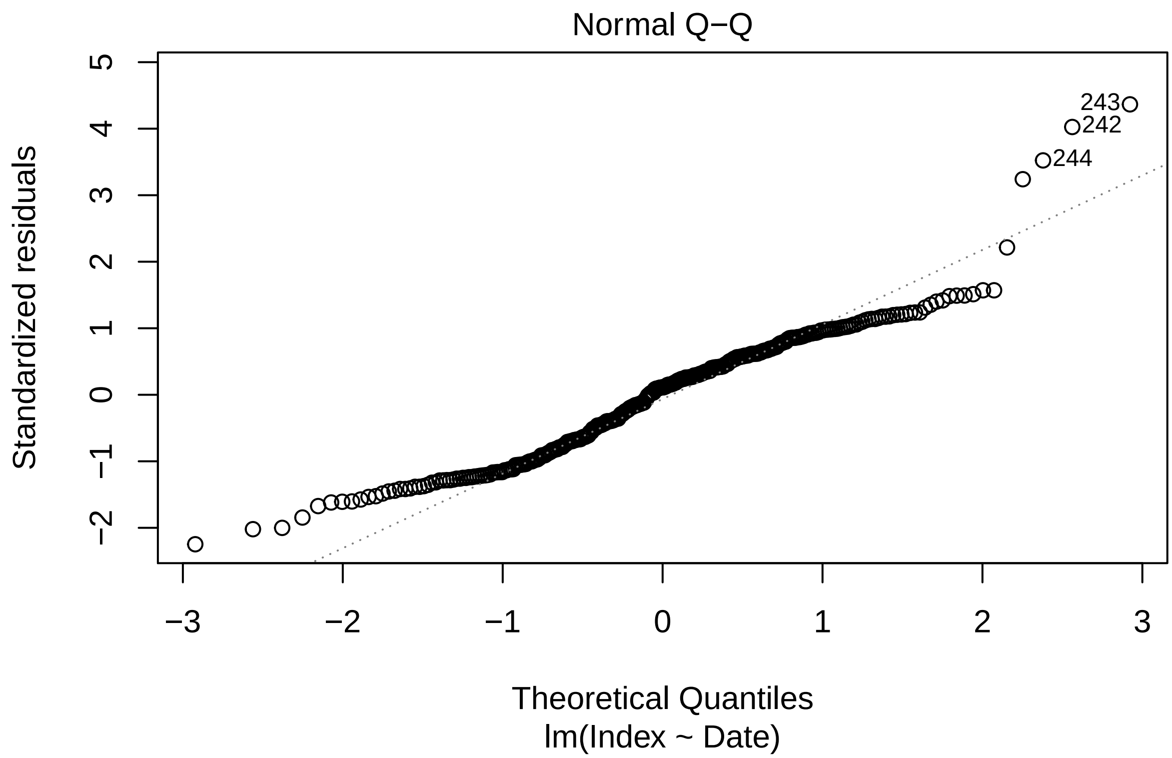
While the normality of the distribution is still questionable and there are a few CPI values that diverge from expectations in our Residuals vs Leverage plot, this data set gives us something we can work with. We recognize that this data is inherently variable and as such may never be truly Bayesian in its distribution. Transformations may be beneficial, in this case, when trying to explain the relationship of CPI over time. To better assess this relationship in an attempt to predict its behavior more accurately, we should split this data into training and testing data sets for model development. We also use a Bayesian classifier to determine the most appropriate transformation on this training data.

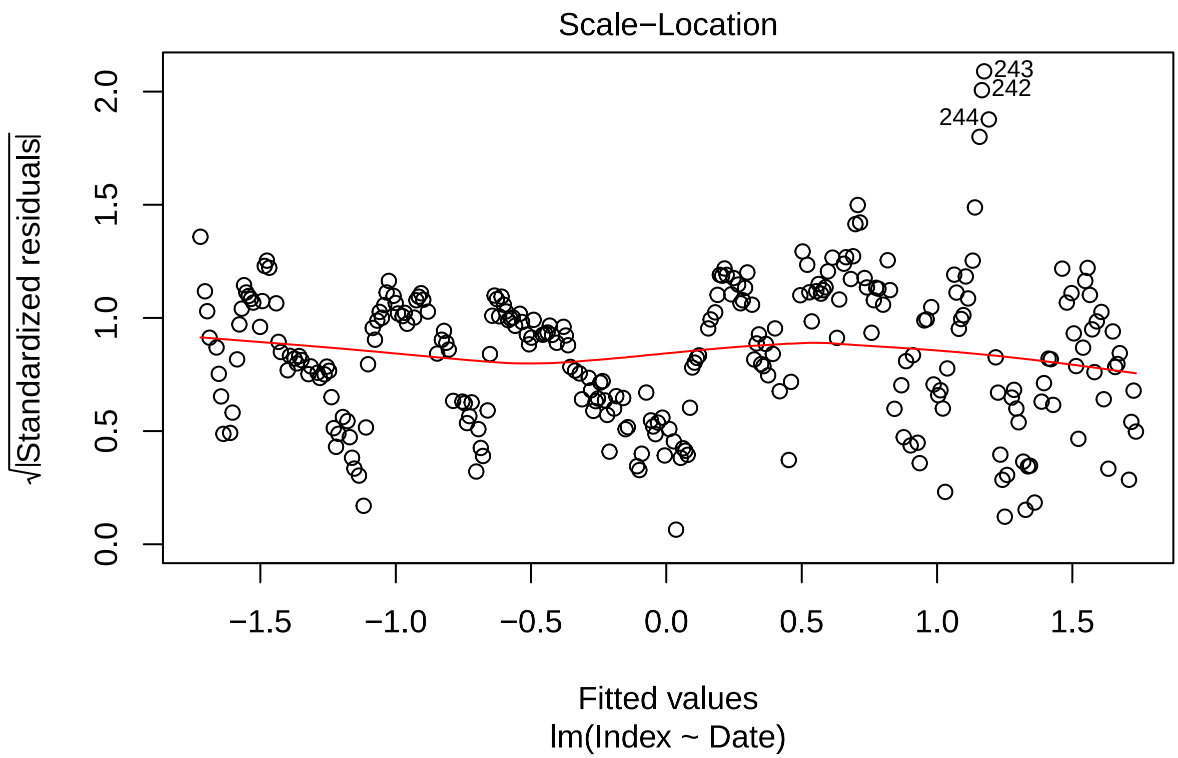
```
## Best Normalizing transformation with 288 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - arcsinh(x): 1.3294
##  - Box-Cox: 1.2658
##  - Center+scale: 1.1924
##  - Exp(x): 34.7778
##  - Log_b(x+a): 1.3294
##  - orderNorm (ORQ): 1.2422
##  - sqrt(x + 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
```

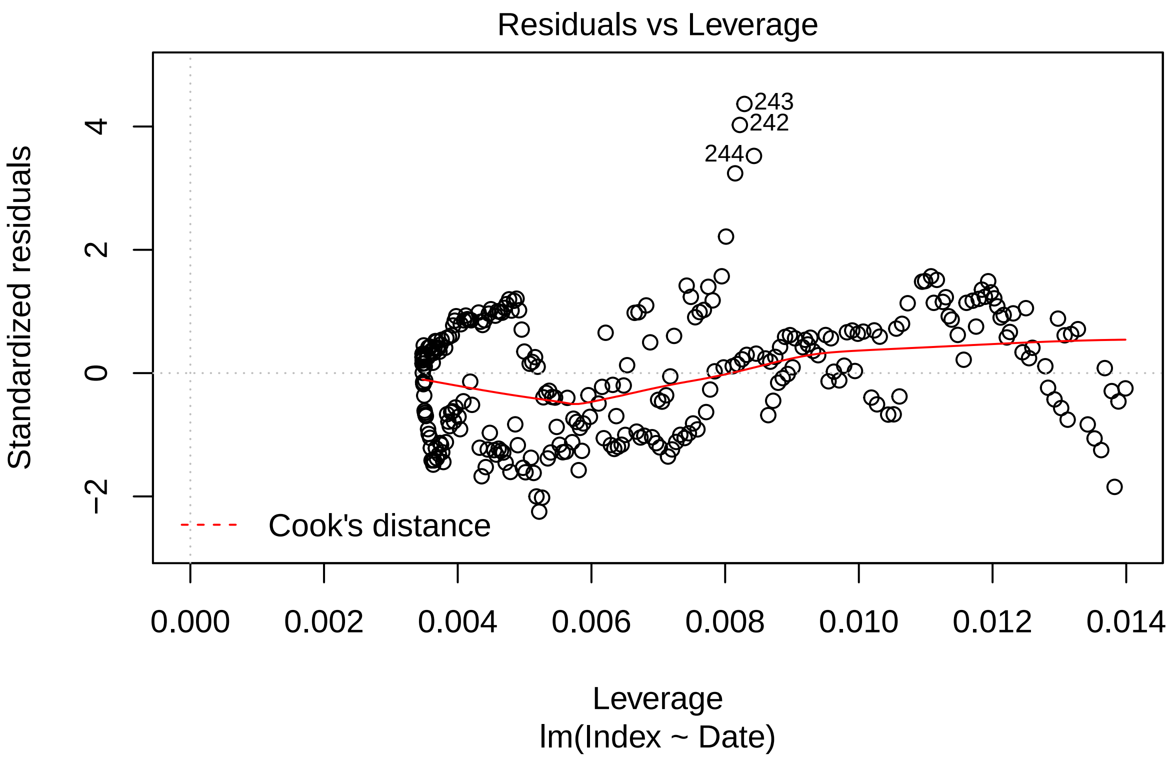
Perhaps ironically based on this classifier, centering and scaling the data is our best option from a host of statistical operations even though it will not adjust normality of residuals. These operations include but are not limited to Box-Cox, Yeo-Johnson, orderNorm, and sqrt which were each close runner-up transformations. We perform these on our newest training data with 70% of the original CPI dates and values. The results are shown:

```
##
## 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
```




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

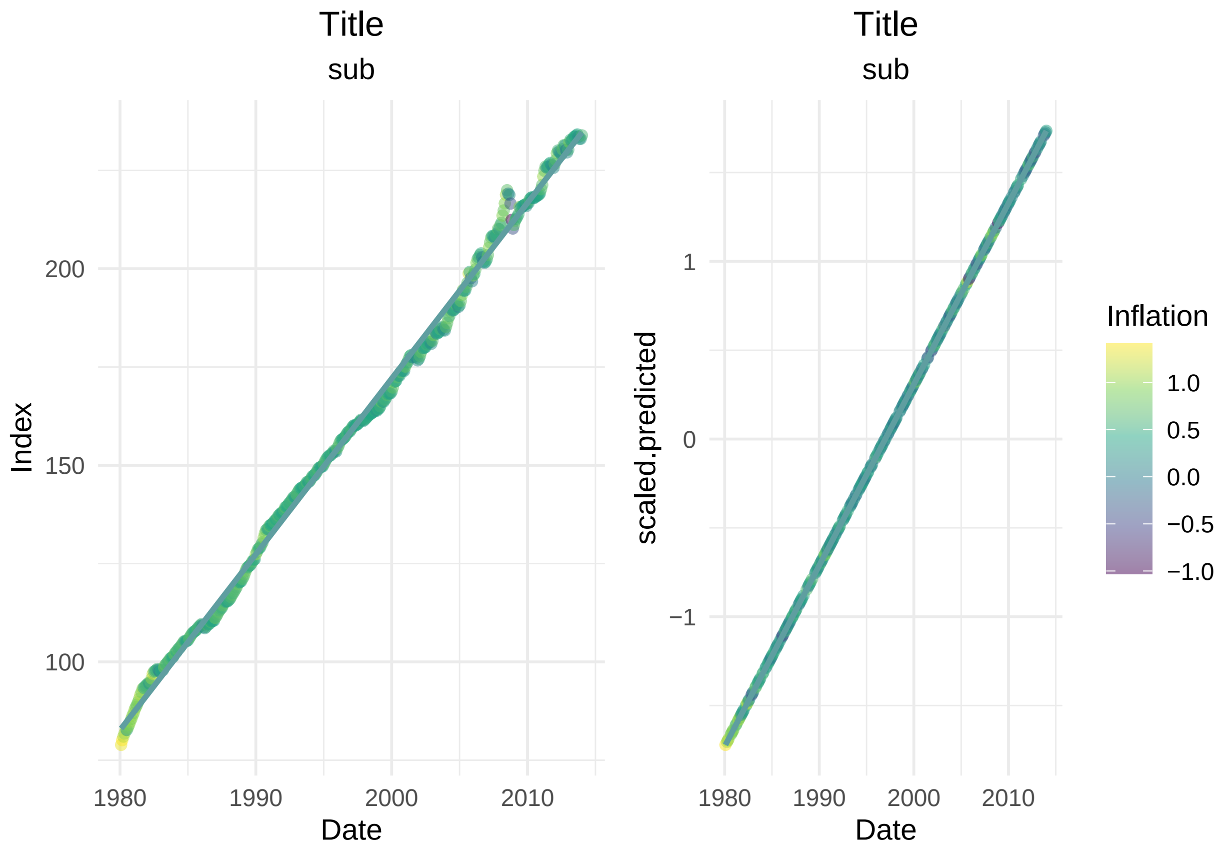

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
243
242
244
Fitted values
lm(Index ~ Date)



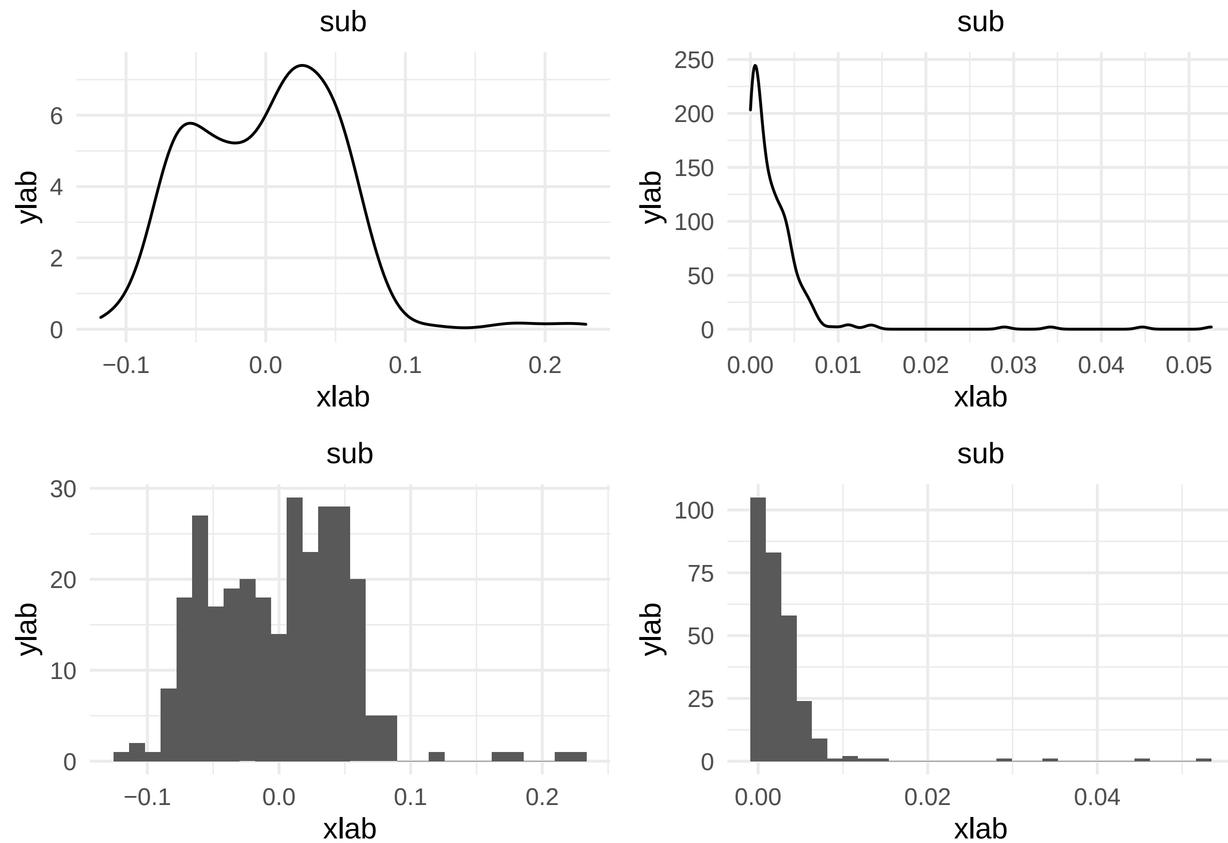
Normality does not change, but the distance of our red trendlines and axis values appear to have shrunk in size, which we suspect benefits the linearity of Residuals vs Fitted as well as our Scale-Location and Residuals vs Leverage plots. At the expense of .1% of our R^2 value, we have successfully reduced the standard error in this model to near 0.05 down from 2.7. We add these statistics to our table and calculate the errors in every prediction, its magnitude, the average error of every prediction, and compute the root mean squared error in subsequent steps.

```
## # A tibble: 288 x 8
##    Date       Index[,1] Inflation scaled.predicted scaled.error[,1]
##    <date>         <dbl>     <dbl>            <dbl>            <dbl>
##  1 1980-02-01     -1.82      1.41            -1.72          -0.0966
##  2 1980-04-01     -1.77      1.12            -1.70          -0.0654
##  3 1980-05-01     -1.75      0.99            -1.70          -0.0555
##  4 1980-06-01     -1.73      1.1             -1.69          -0.0436
##  5 1980-09-01     -1.70      0.84            -1.66          -0.0396
##  6 1980-10-01     -1.68      0.95            -1.65          -0.0297
##  7 1980-11-01     -1.67      0.83            -1.64          -0.0224
##  8 1980-12-01     -1.65      0.94            -1.64          -0.0125
##  9 1981-03-01     -1.60      0.68            -1.61           0.0127
## 10 1981-04-01     -1.59      0.68            -1.60           0.0177
## # ... with 278 more rows, and 3 more variables: scaled.mag[,1] <dbl>,
## #   scaled.eravg <dbl>, scaled.rmse <dbl>
```

Our resultant model produces values that fall almost exactly on the regression line, indicating a near perfect fit for an imperfect data set. Inflation values are again highlighted for reference on a continuous scale. Notice how the points of change in inflation did not change, only the location of points on the plot and the scale of the axes. Over the same time frame, our scaled prediction values perform marginally better than the original contemporary model.



As a final check, we review the standard errors of each prediction and its magnitude as both density plots and histograms. At this scale, nearly all errors occur between -0.1 and 0.1 with the magnitude of those errors rarely ever having an occurrence greater than .01. This is good news for the contemporary model hypothesis.



Discussion and Conclusion

Given recent citations in literature about the significant level of new stability in inflation measures we hypothesized that CPI data from the first of January in 1980 through the latest adjusted points we could find in 2014 would show that a contemporary model is justified for predicting CPI and its related affiliate statistics. In our experiment the old-fashioned model failed all but one of the assumptions of linear regression (leverage remained) and was only able to explain about 63.8% of the data with a t-statistic of 79.69 on 563 degrees of freedom (DF) and was statistically significant beyond an alpha level of 0.001. Meanwhile, our contemporary model explained 99.7% of the data, with a p-value less than 0.001 and a t-statistic of 72.95 with 407 DF. Lastly, after the data underwent a 70-30 split with a center-scale transformation, our scaled contemporary model improved upon these scores by reducing the standard error of the contemporary model to 0.053 on 286 degrees of freedom while all other test statistics remained nearly unchanged. Our final root mean squared error for this scaled contemporary model was 0.0525, with an mean error of 0.0028, and no error larger than 0.229. Based on this experiment modeling the segmented time periods, there is evidence to support making predictions on the basis that data prior to 1980 is impractical for prediction in modern times.

References

[1]Baker, E. S. (2014, May). Modeling core inflation: Considering goods and services separately. Retrieved from https://www.bls.gov/opub/mlr/2014/beyond-bls/pdf/modeling-core-inflation-considering-goods-and-services-separately.pdf [2]Board of Governors, F. (2021, May 03). Effective federal funds rate. Retrieved from https://fred.stlouisfed.org/series/FEDFUNDS [3]Brissimis, S. N., Economic

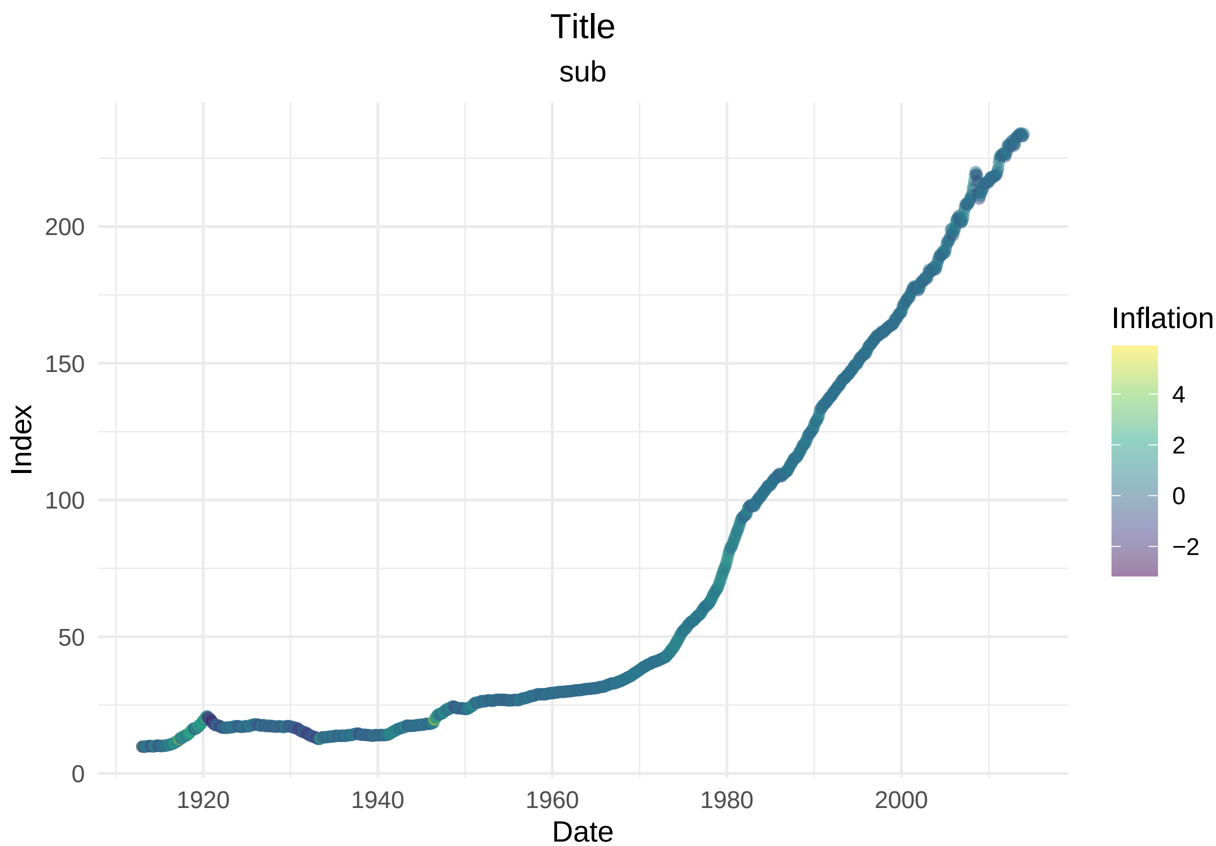
Research Department, Bank of Greece, & Magginas, N. S., Department of Economics, University of Piraeus. (2008). Inflation Forecasts and the New Keynesian Phillips Curve. International Journal of Central Banking, (13). doi:https://www.ijcb.org/journal/ijcb08q2a1.pdf [4]Bureau of Labor Statistics. (2021, May 12). Archived consumer price index supplemental files. Retrieved from https://www.bls.gov/cpi/tables/supplemental-files/home.htm [5]Chao, E. L., U.S. Department of Labor, & Utgoff, K. P., U.S. Bureau of Labor Statistics. (2006, May). 100 Years of U.S. Consumer Spending: Data for the Nation, New York City, and Boston. Retrieved from https://www.bls.gov/opub/100-years-of-u-s-consumer-spending.pdf [6]Chien, N., & Mistry, R. (2013). Geographic variations in cost of living: Associations with family and child well-being. Retrieved May 22, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3509251/ [7]Consumer price index frequently asked questions. (2021). Retrieved from https://www.bls.gov/cpi/questions-and-answers.htm [8]Data source from data.io: Consumer Price Index for All Urban Consumers (CPI-U) from U.S. Department Of Labor Bureau of Labor Statistics. Retrieved from using online link: https://pkgstore.datahub.io/core/cpi-us/8/datapackage.json [9]DeSilver, D. (2020, May 30). How America's diet has changed over time. Retrieved from https://www.pewresearch.org/fact-tank/2016/12/13/whats-on-your-table-how-americas-diet-has-changed-over-the-decades/ [10]Fernando, J. (2021, May 20). Consumer price index (cpi) definition. Retrieved from https://www.investopedia.com/terms/c/consumerpriceindex.asp#:~:text=The%20CPI%20measures%20the%20average,services%2C%20commonly%20known%20as%20inflation.&text=T [11]James H. Stock & Mark W. Watson, 2008. "Phillips curve inflation forecasts," Conference Series ; [Proceedings], Federal Reserve Bank of Boston, vol. 53. [12]Jolliffe, D. (2006, November 01). Adjusting for living costs can change who is considered poor. Retrieved from https://www.ers.usda.gov/amber-waves/2006/november/adjusting-for-living-costs-can-change-who-is-considered-poor/ [13]Jolliffe, D. (2006, November 01). Adjusting for Living Costs Can Change Who Is Considered Poor. Retrieved May 21, 2021, from https://www.ers.usda.gov/amber-waves/2006/november/adjusting-for-living-costs-can-change-who-is-considered-poor/ [14]Jørgensen, Peter Lihn, Kevin J. Lansing. 2021. "Anchored Inflation Expectations and the Flatter Phillips Curve," Federal Reserve Bank of San Francisco Working Paper 2019-27. https://doi.org/10.24148/wp2019-27 [15]Koba, M. (2013, May 01). Consumer price Index: CNBC Explains. Retrieved from https://www.cnbc.com/id/43769766 [16]Peach, R., Rich, R., & Linder, M. H. (2013). The Parts Are More Than the Whole: Separating Goods and Services to Predict Core Inflation. Current Issues in Economics and Finance, 19(7). doi:https://www.newyorkfed.org/medialibrary/media/research/current_issues/ci19-7.pdf [17]Staff. Bureau of Labor Statistics. (2021, May 19). Consumer price index up 4.2 percent from April 2020 to April 2021. Retrieved from https://www.bls.gov/opub/ted/2021/consumer-price-index-up-4-2-percent-from-april-2020-to-april-2021.htm [18]Stock, J. H., & Watson, M. W. (2008). Phillips curve inflation forecasts. National Bureau of Economic Research, 53, 14322nd ser. doi:10.3386/w14322 [19]Stock, J. H., & Watson, M. W. (2019). Forecasting Inflation. Retrieved from https://scholar.harvard.edu/files/stock/files/forecastinginflation.pdf

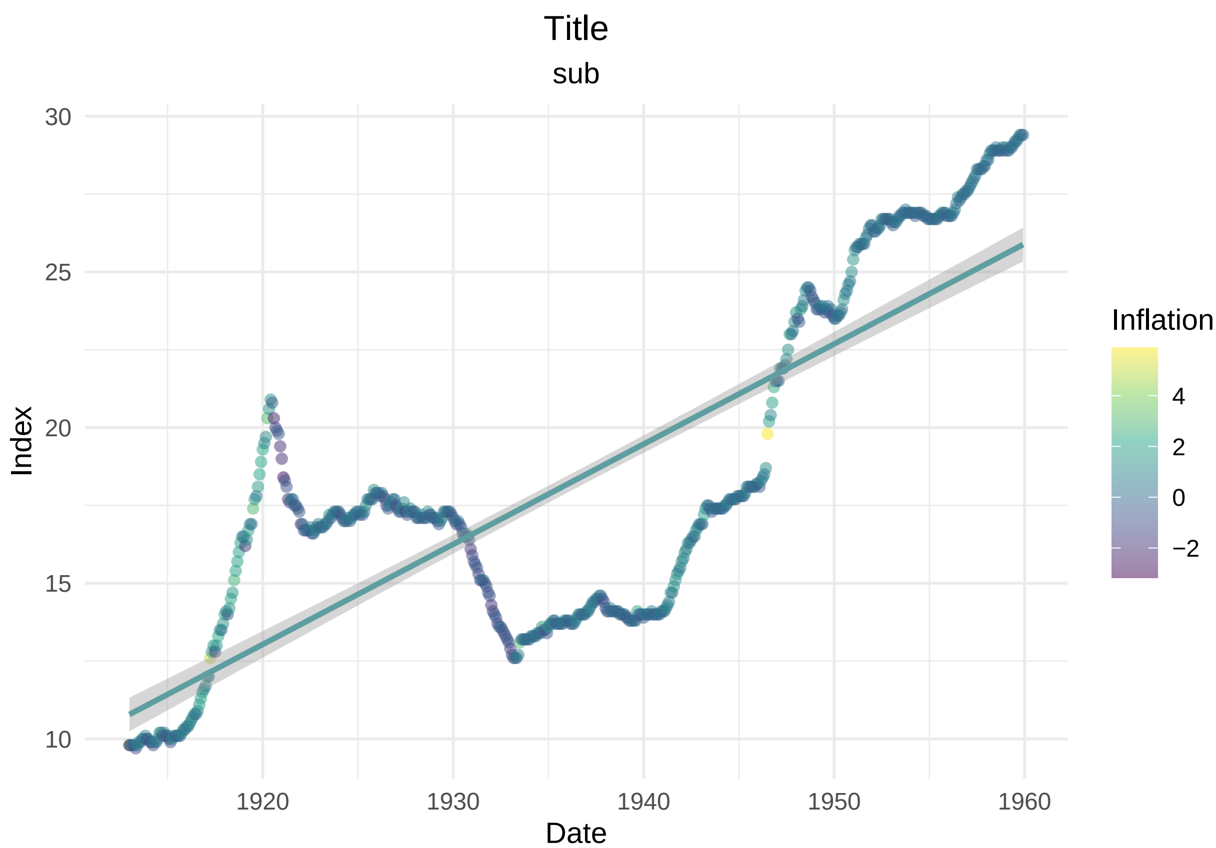
Appendecies

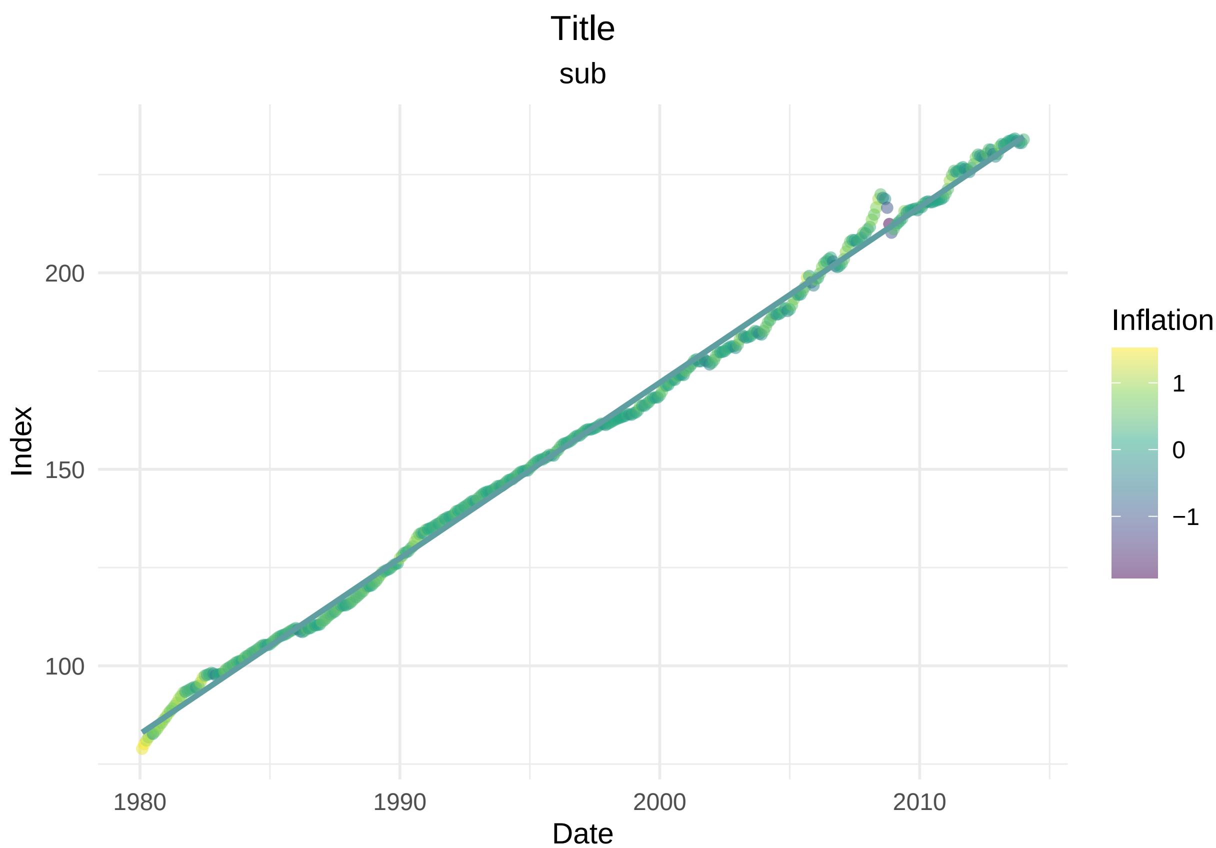
```
## # A tibble: 6 x 3
##   Date       Index Inflation
##   <date>     <dbl>     <dbl>
## 1 1913-01-01   9.8     NA
## 2 1913-02-01   9.8      0
## 3 1913-03-01   9.8      0
## 4 1913-04-01   9.8      0
## 5 1913-05-01   9.7     -1.02
## 6 1913-06-01   9.8      1.03
```

```
cpi %>%
  ggplot(aes(Date, Index)) + geom_point(aes(color=Inflation)) +
  scale_color_viridis_c(aesthetics = c("colour", "fill"), option = "D", alpha = .5) +
```

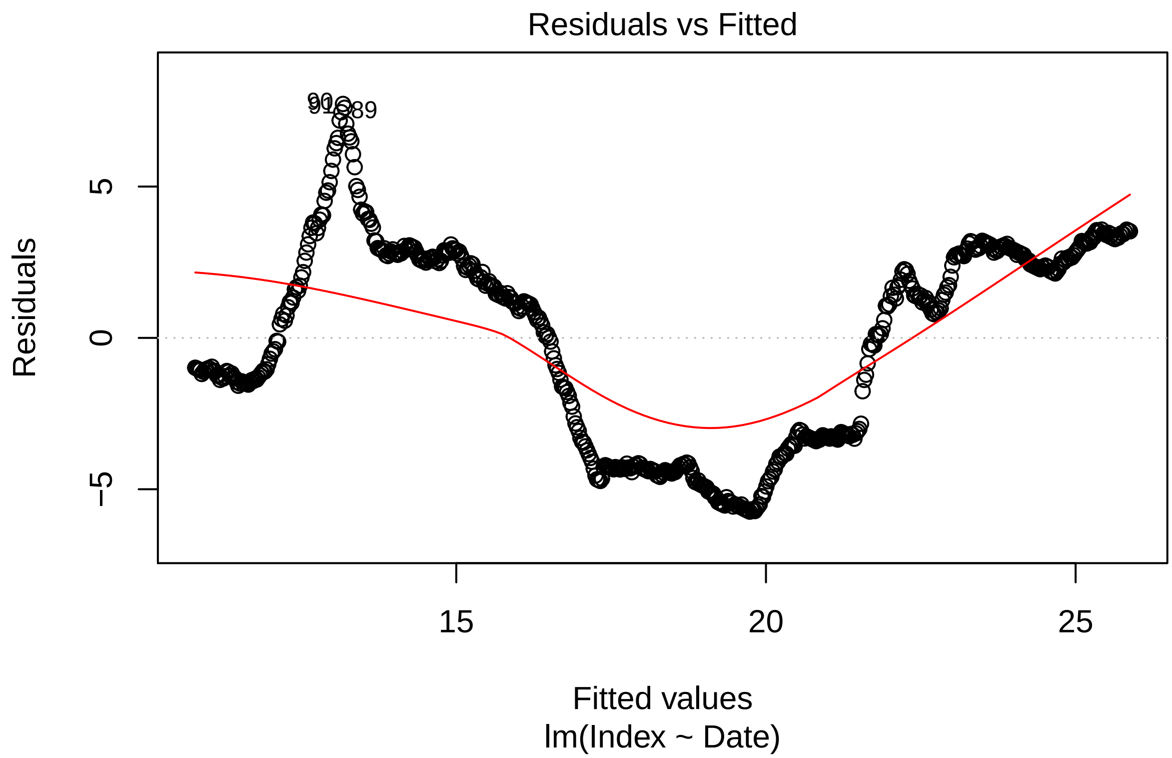
```
labs(title = "Title", subtitle = "sub", xlab="xlab", ylab="ylab") +
theme(plot.title = element_text(hjust = .5), plot.subtitle = element_text(hjust = .5))
```

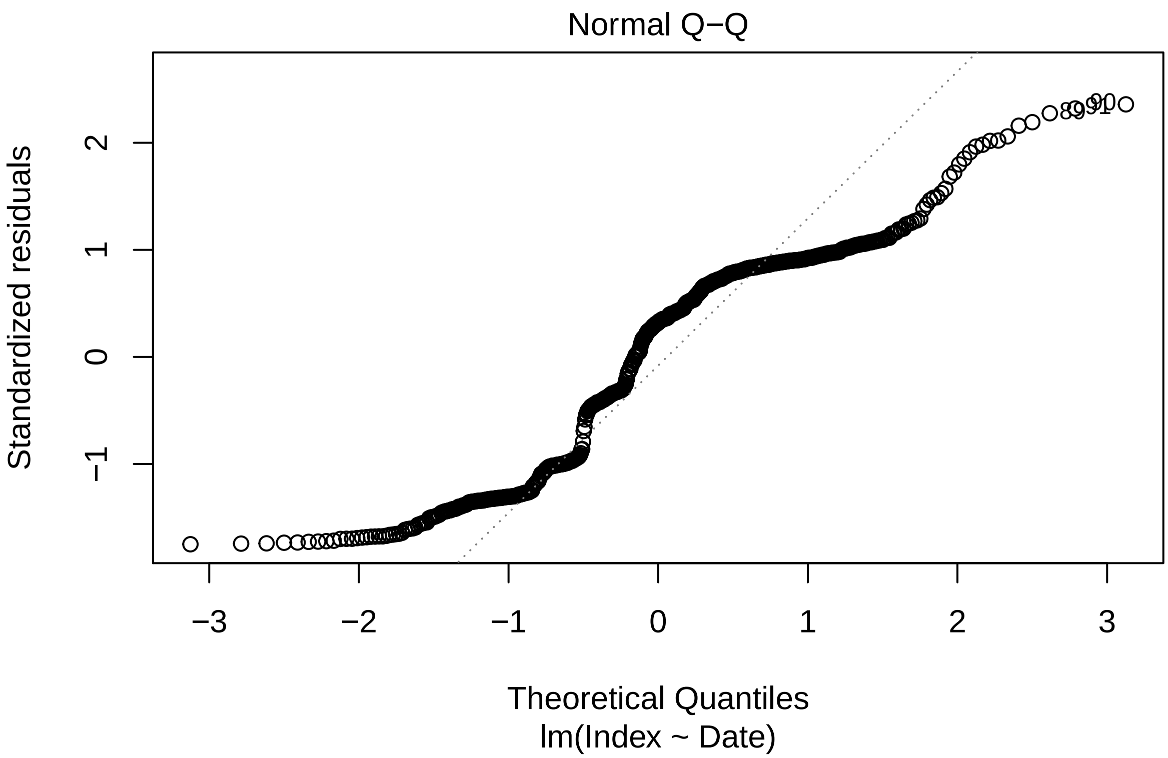


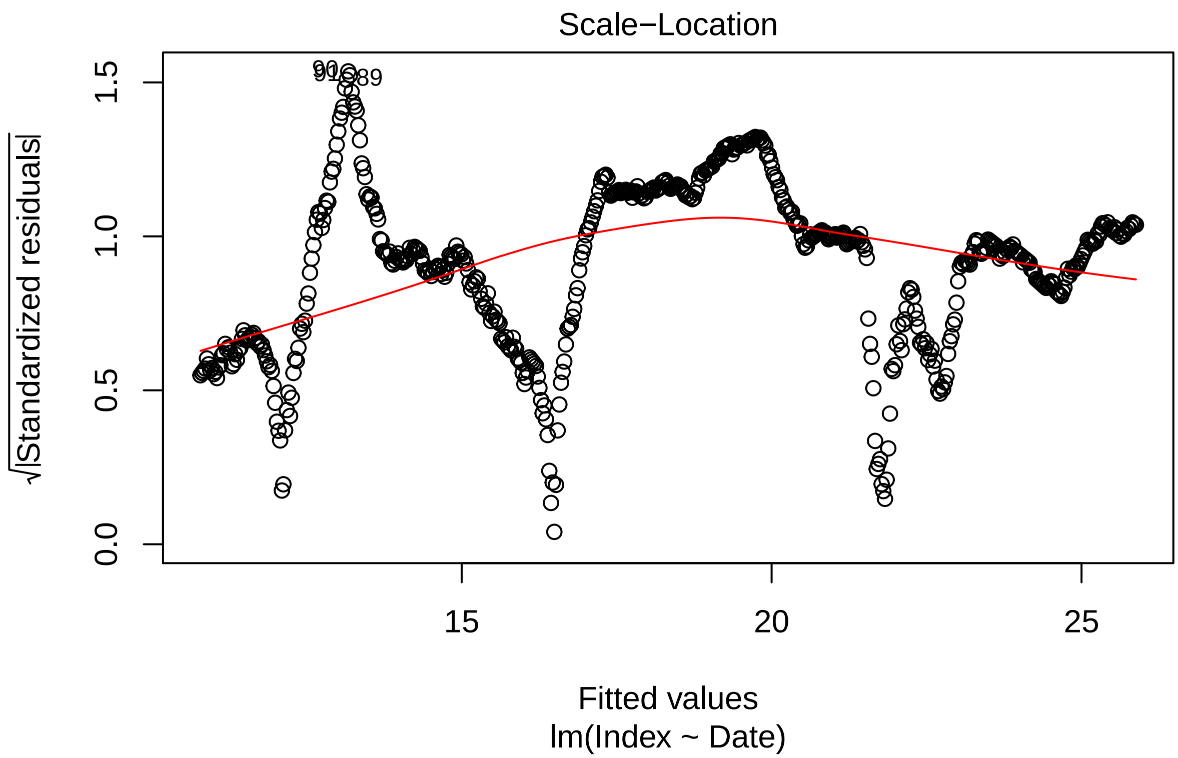

Title
sub
Index
Date
Inflation
30
25
20
15
10
1920
1930
1940
1950
1960
4
2
0
−2



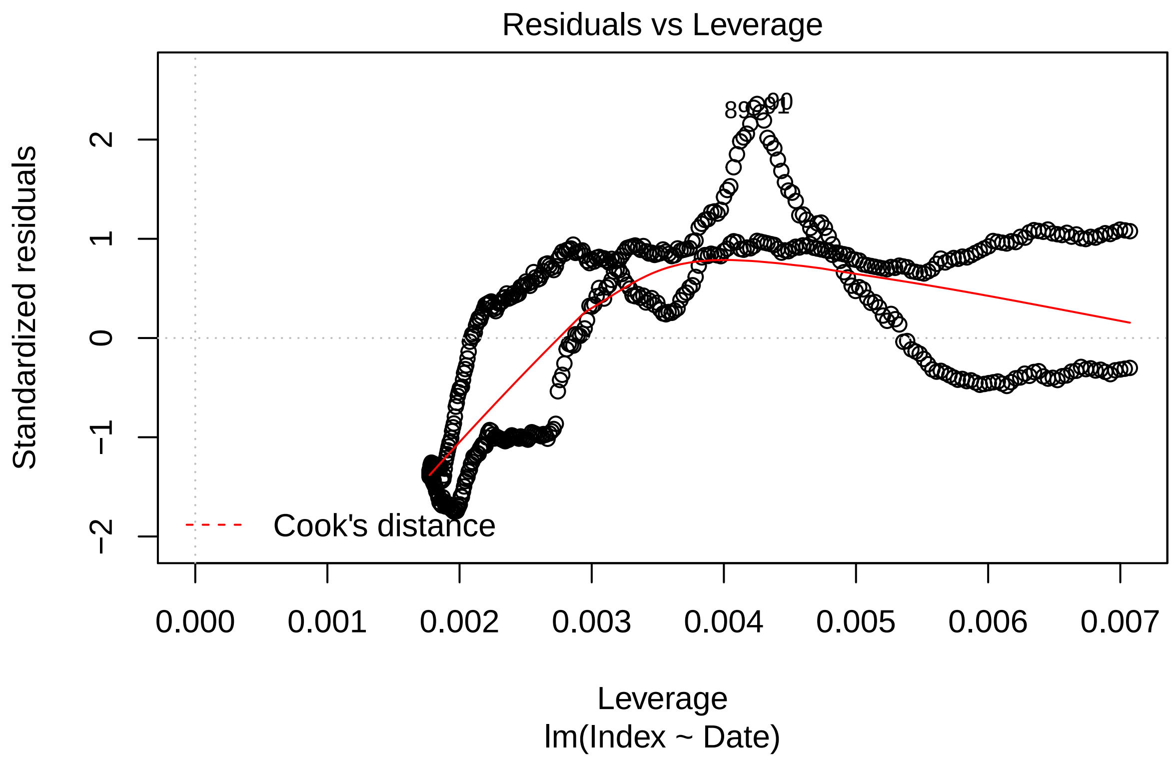

```
##
## Call:
## lm(formula = Index ~ Date, data = g1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.739 -3.305  1.050  2.773  7.729
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.912e+01  3.687e-01   78.98   <2e-16 ***
## Date        8.808e-04  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
```


Residuals vs Fitted
Residuals
Fitted values
lm(Index ~ Date)
5
0
−5
15
20
25
90
91
89

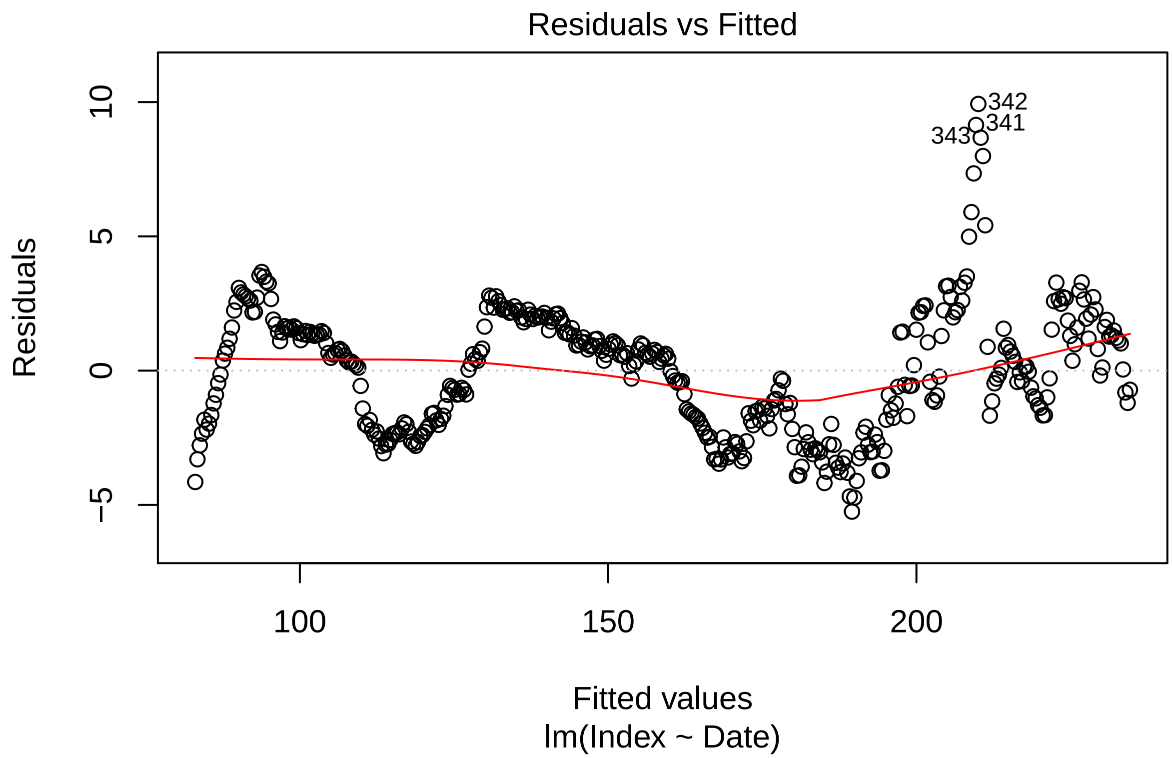

Normal Q−Q
Standardized residuals
−1
0
1
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Index ~ Date)



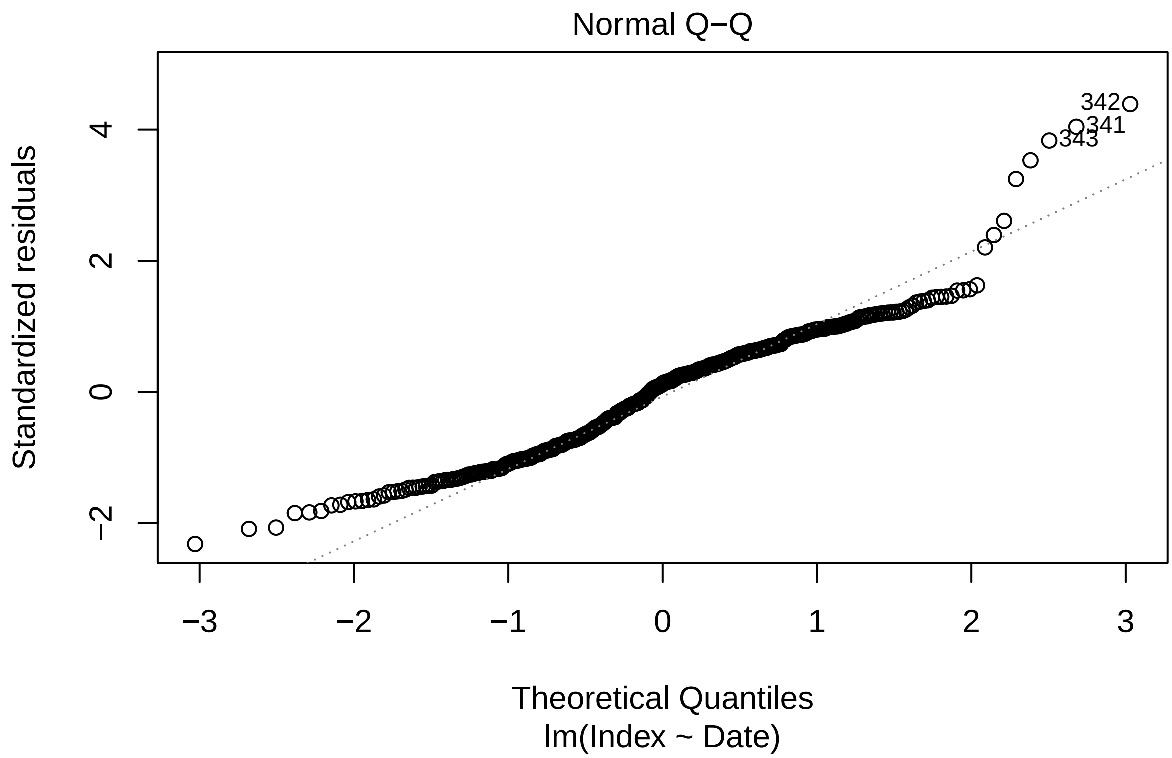
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
Fitted values
lm(Index ~ Date)
90
91
89

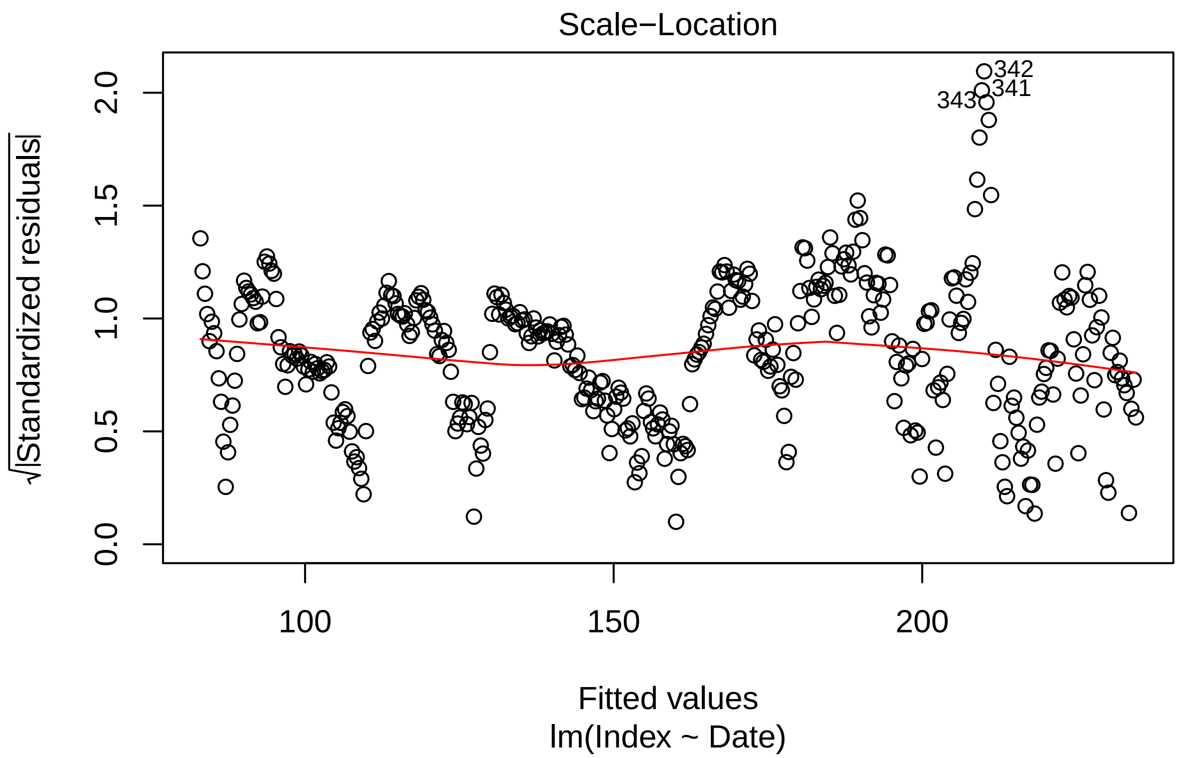



```
##
## Call:
## lm(formula = Index ~ Date, data = g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2504 -1.8410  0.2805  1.5275  9.9303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.798e+01  3.293e-01   115.3   <2e-16 ***
## Date        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
```

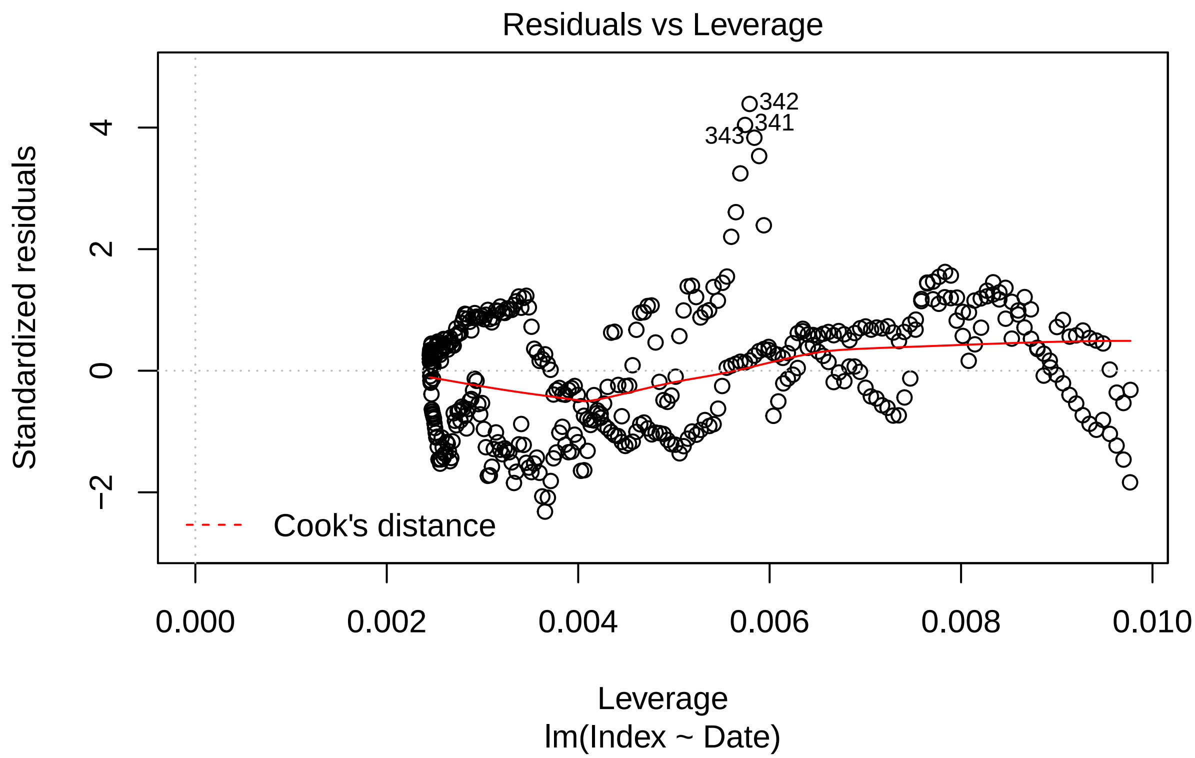


Residuals vs Fitted
Residuals
Fitted values
lm(Index ~ Date)
10
5
0
−5
100
150
200
342
341
343


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



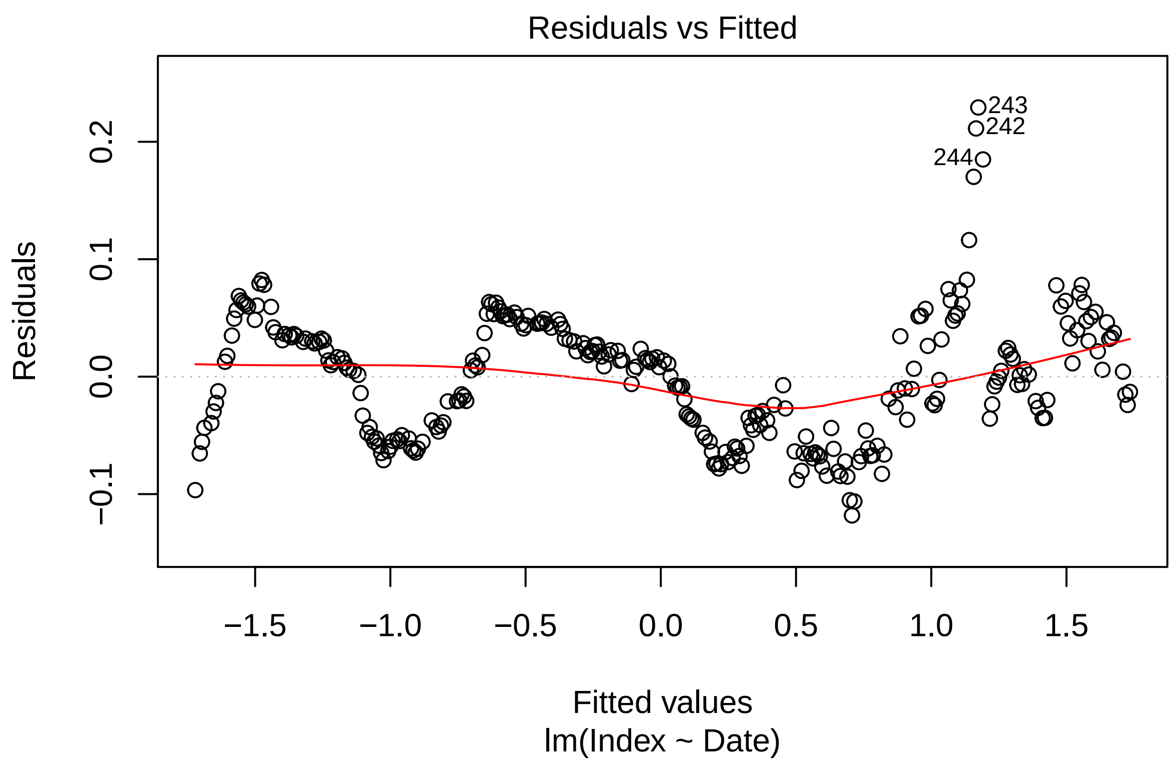
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
150
200
Fitted values
lm(Index ~ Date)
342
341
343

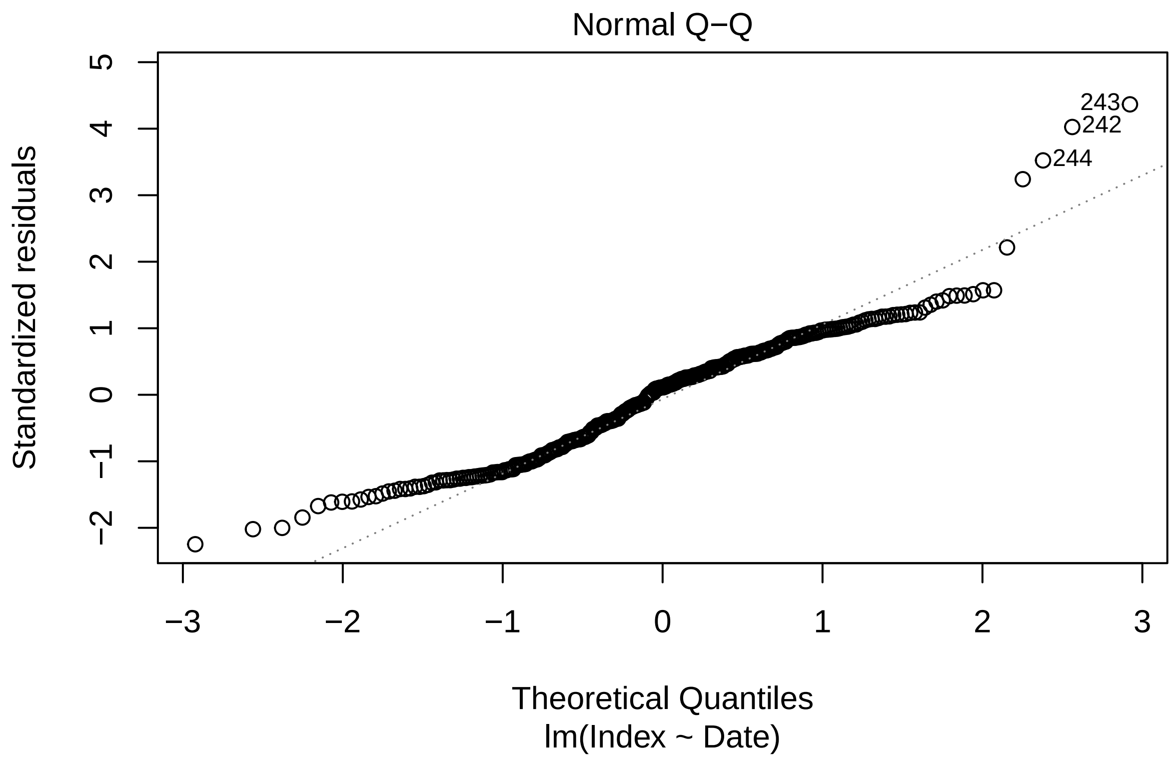


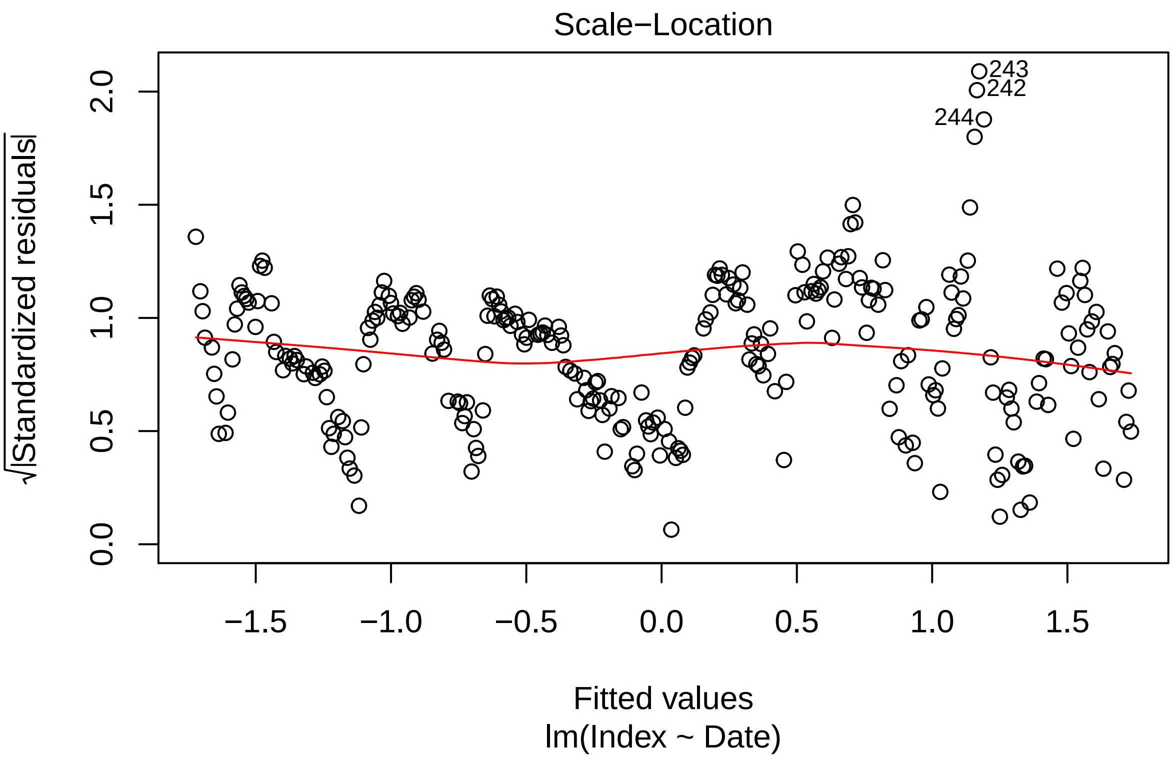

```
## Best Normalizing transformation with 288 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - arcsinh(x): 1.3294
##  - Box-Cox: 1.2658
##  - Center+scale: 1.1924
##  - Exp(x): 34.7778
##  - Log_b(x+a): 1.3294
##  - orderNorm (ORQ): 1.2422
##  - sqrt(x + 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
```

```
##
## Call:
## lm(formula = Index ~ Date, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118151 -0.043081  0.005851  0.036281  0.229216
```

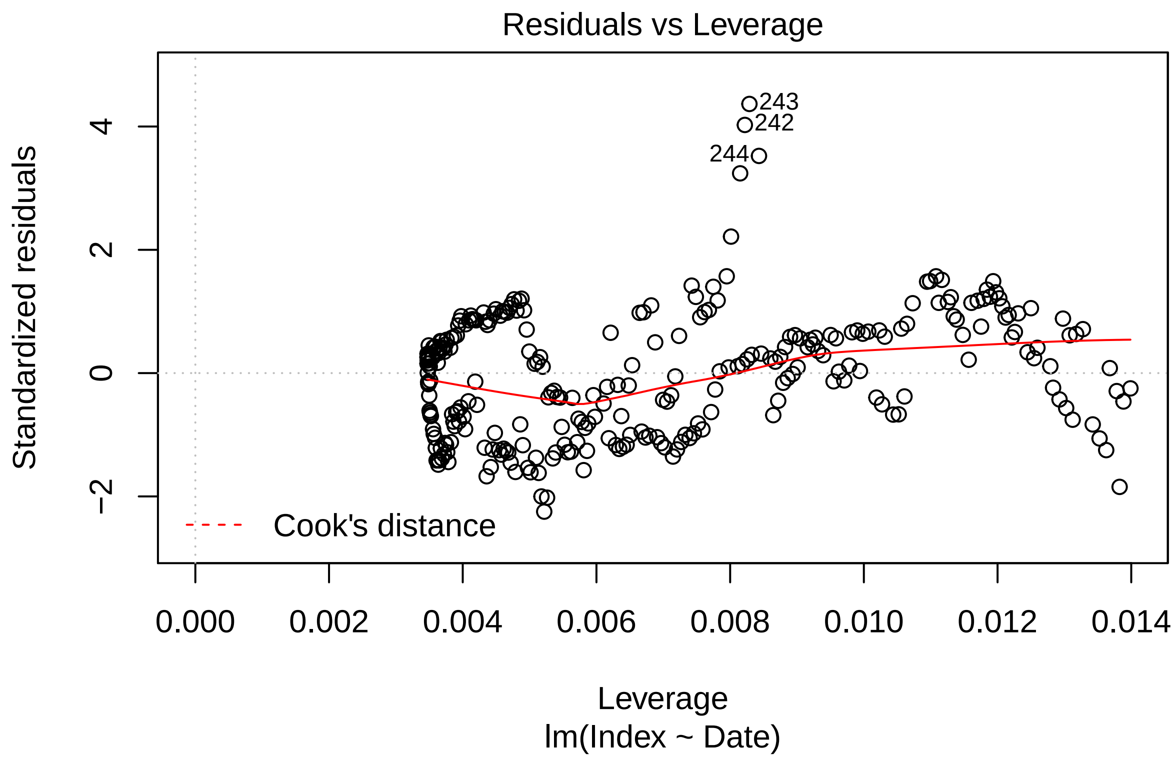
```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.749e+00  9.113e-03  -301.7   <2e-16 ***
## Date         2.790e-04  8.695e-07   320.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05272 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
```



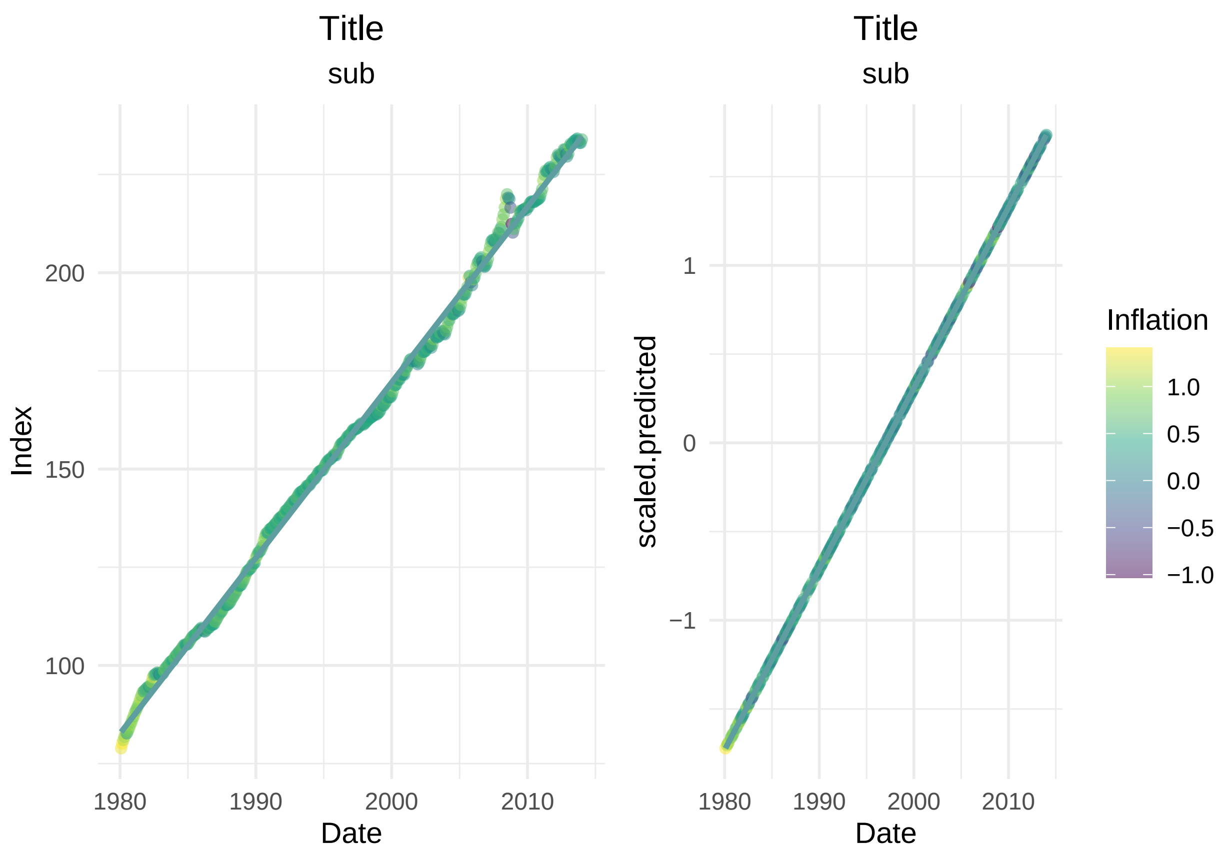

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Index ~ Date)
243
242
244
−3
−2
−1
0
1
2
3
4
5



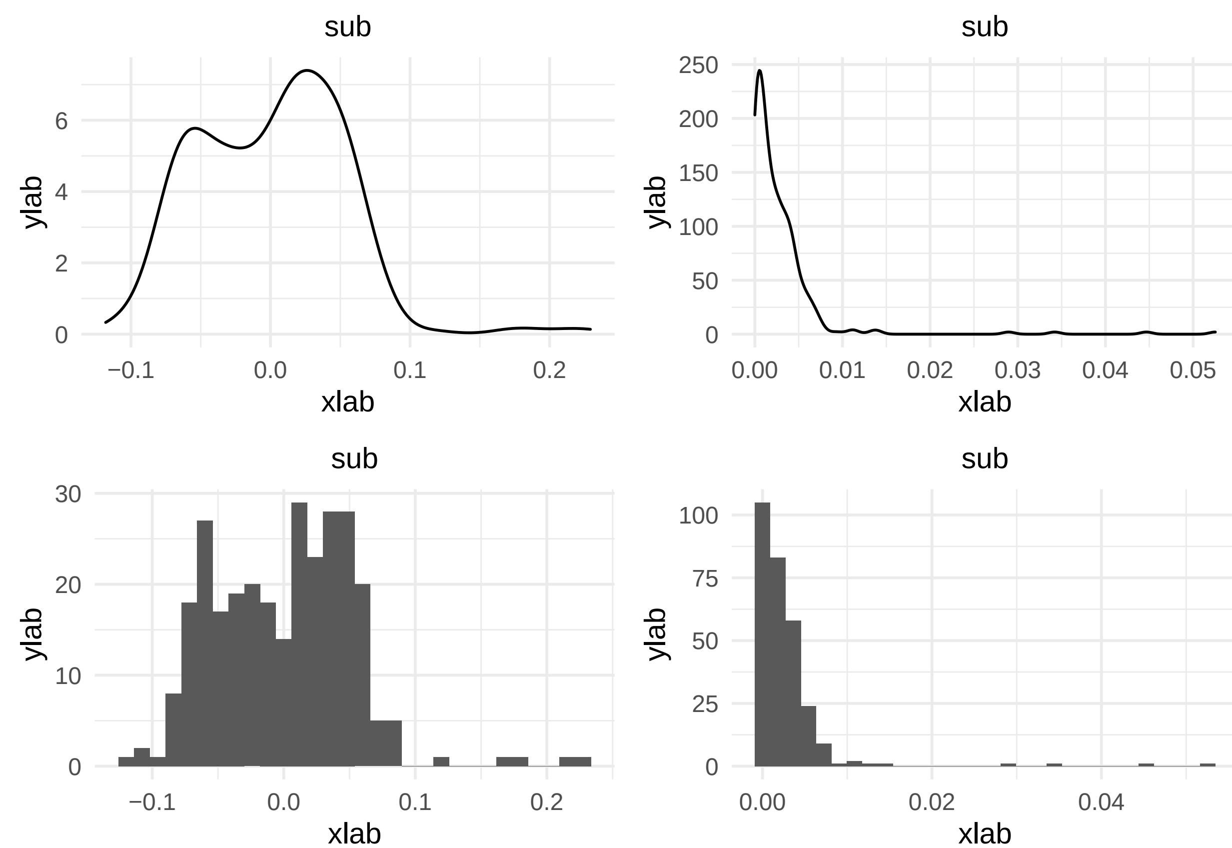
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
243
242
244
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Fitted values
lm(Index ~ Date)




```
## # A tibble: 288 x 8
##    Date       Index[,1] Inflation scaled.predicted scaled.error[,1]
##    <date>         <dbl>     <dbl>            <dbl>            <dbl>
##  1 1980-02-01     -1.82      1.41            -1.72          -0.0966
##  2 1980-04-01     -1.77      1.12            -1.70          -0.0654
##  3 1980-05-01     -1.75      0.99            -1.70          -0.0555
##  4 1980-06-01     -1.73      1.1             -1.69          -0.0436
##  5 1980-09-01     -1.70      0.84            -1.66          -0.0396
##  6 1980-10-01     -1.68      0.95            -1.65          -0.0297
##  7 1980-11-01     -1.67      0.83            -1.64          -0.0224
##  8 1980-12-01     -1.65      0.94            -1.64          -0.0125
##  9 1981-03-01     -1.60      0.68            -1.61           0.0127
## 10 1981-04-01     -1.59      0.68            -1.60           0.0177
## # ... with 278 more rows, and 3 more variables: scaled.mag[,1] <dbl>,
## #   scaled.eravg <dbl>, scaled.rmse <dbl>
```



Title
sub
Index
200
150
100
1980
1990
2000
2010
Date
Title
sub
scaled.predicted
1
0
−1
1980
1990
2000
2010
Date
Inflation
1.0
0.5
0.0
−0.5
−1.0




```
## [1] 0.2292165

##
##  One Sample t-test
##
## data:  g1$Index
## t = 79.685, df = 563, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  17.87935 18.78306
## sample estimates:
## mean of x
##  18.33121

##
##  One Sample t-test
##
## data:  g2$Index
## t = 72.949, df = 407, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  154.5473 163.1073
## sample estimates:
## mean of x
##  158.8273
```