Time Series Analysis of Sales Growth

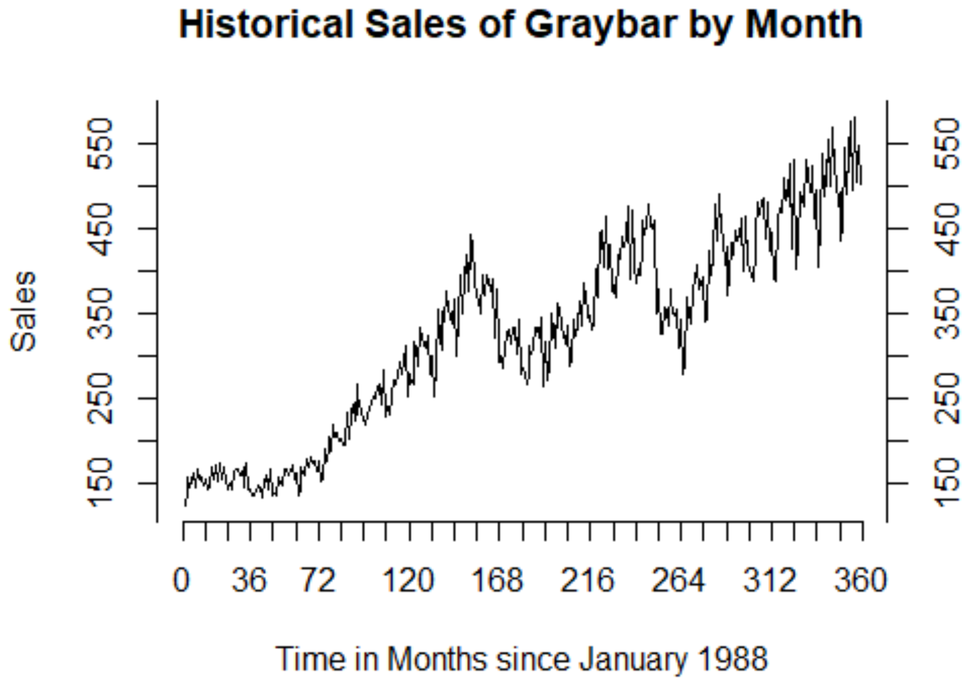
Josh Liu

January 4, 2018

Introduction

In this analysis, I will set up two models. In the first model, I will take a look at the Company's sales data, without taking into consideration external macroeconomic indicators. In the second model, I will use all available variables, including sales data and external macroeconomic indicators.

The data being used were supplied to me on January 3, 2018, by Dennis. The data date back to January 1988, through November 2017.

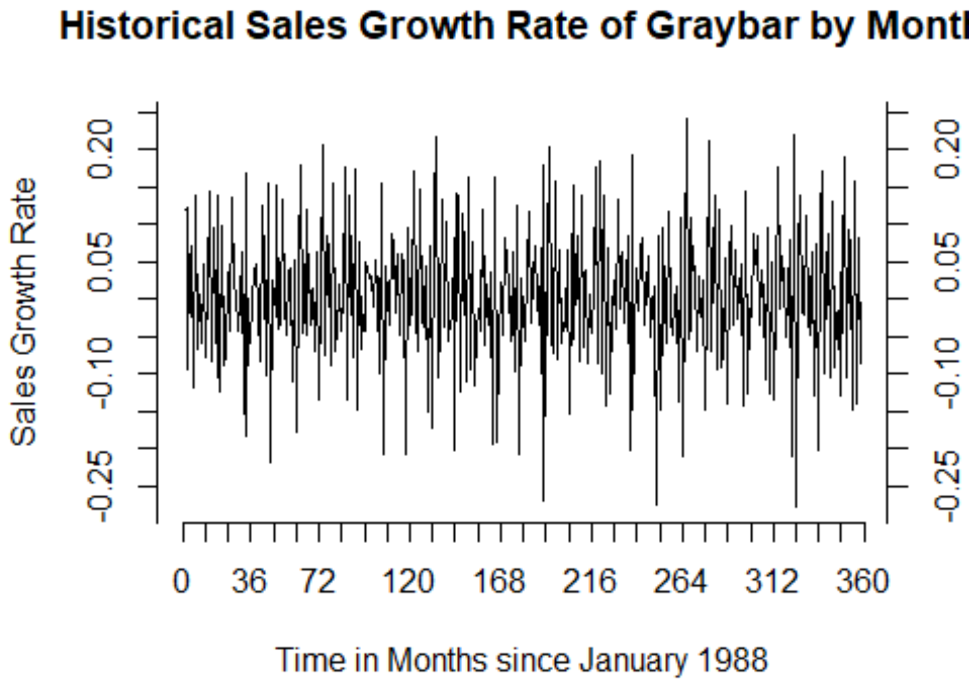


Clearly, sales have been growing since January 1988. There were up-and-downs, but the general pattern is moving upward. Next, I will try to see if it is possible to identify a periodicity of the pattern of monthly sales.

First, monthly sales are to be transformed into sales growth rate. The growth rate during the T-th month is given by

$$r = \ln \frac{sales(T)}{sales(T-1)},$$

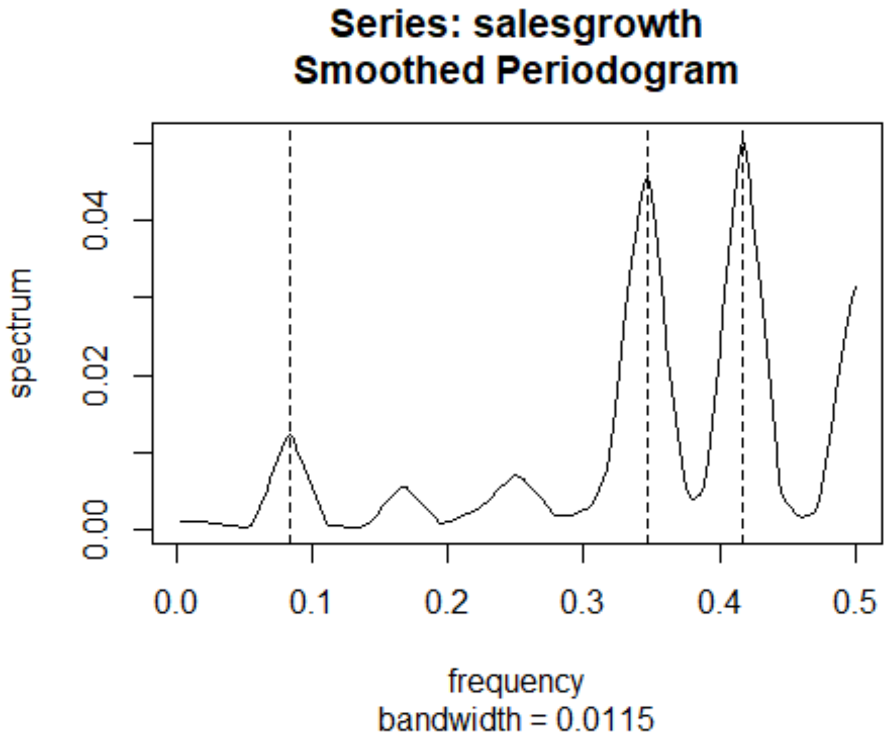
where $sales(T)$ represents the sales of the T-th month, and ln represents natural logarithm. Note that after transformation, we have the growth rate from February 1988, to November 2017, one less data point than the original data on monthly sales.



The plot of growth rate shows a high degree of oscillation. It is my speculation that if we learn how the growth rate oscillates, we may be able to know more about the business/sales cycle, and make more accurate prediction on future sales.

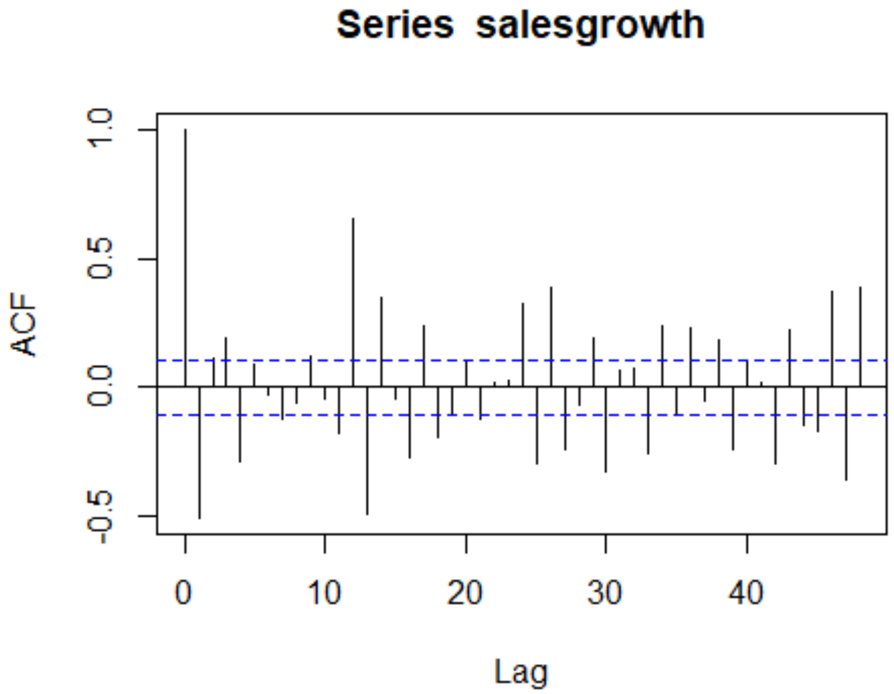
Spectral Analysis

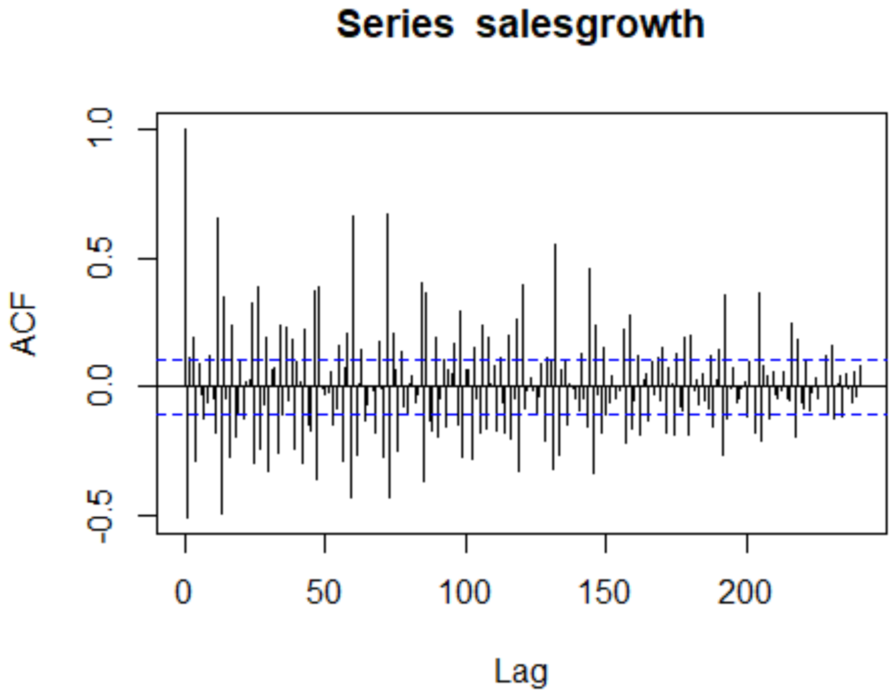
Now, I will analyze the sales growth rate with a technique called "spectral analysis". Spectral analysis works like this: it takes a series of time-indexed data as input, applies discrete Fourier transform to these data points, and outputt the weight/density of the frequency of oscillation. It might sound obscure, but this graph will make it more clear:

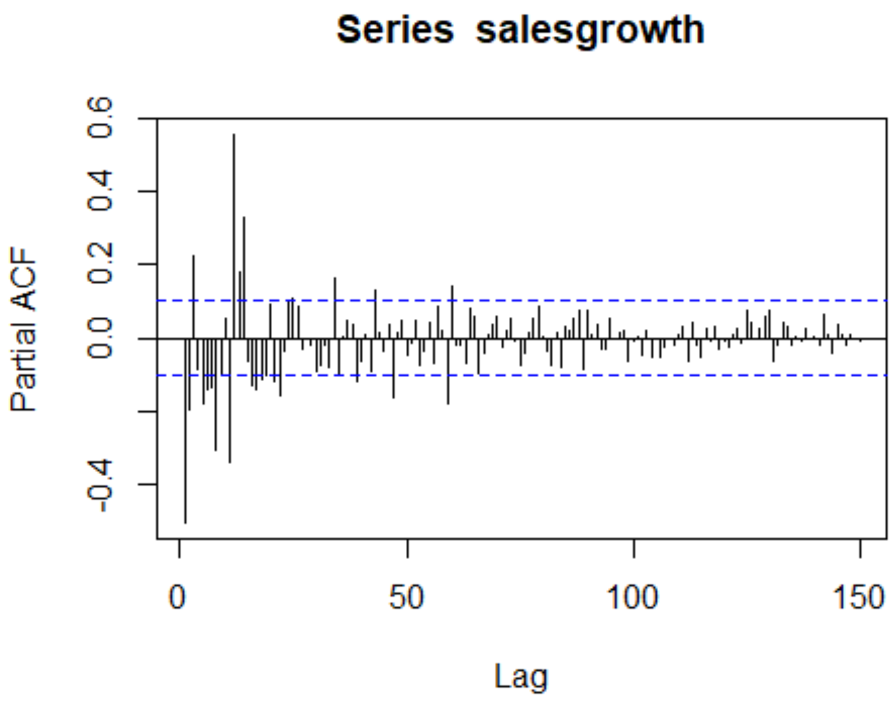


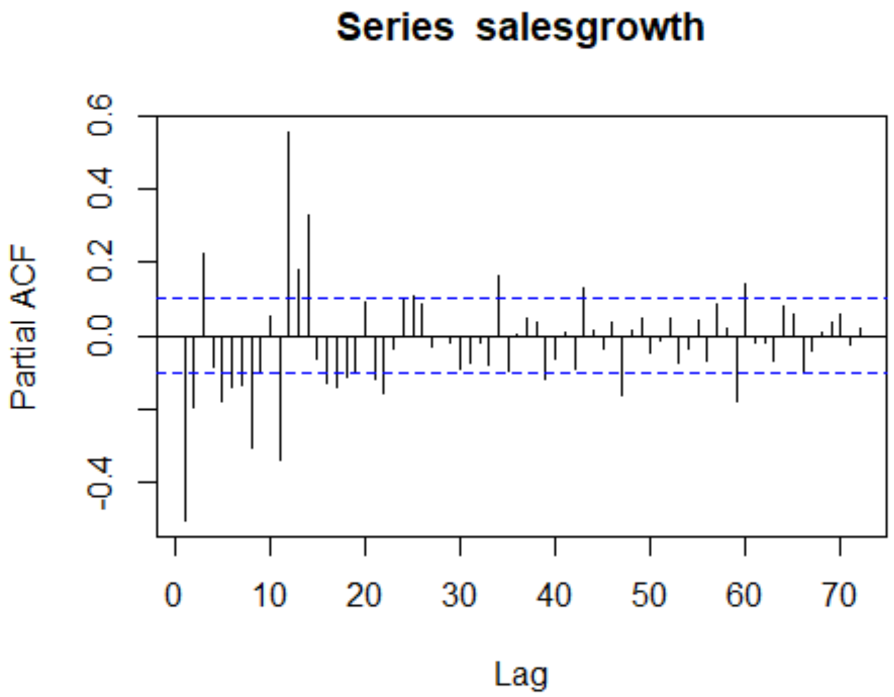
This graph, called “smoothed periodogram,” shows that the maximum spectrums are at the frequencies of 0.42, 0.35, and 0.83 (rounded). Therefore, the most prominent periods of the sales growth rate are 2.4 , 2.88, and 12 (the reciprocal of frequencies). That is to say, sales growth oscillate with the periods 12 months and 3 months.

Now, let’s take a look at the sample auto-correlation of growth rate:


Series salesgrowth
ACF
1.0
0.5
0.0
-0.5
0
10
20
30
40
Lag


Series salesgrowth
ACF
1.0
0.5
0.0
-0.5
0
50
100
150
200
Lag





The autocorrelation chart shows significant correlation at the lags of 1 month, the multiples of 12 months, and $12k \pm 1$ month where k is integer. Looking at the autocorrelation chart on a larger range of lags, we notice that it never decays to zero. At the same time, the partial-autocorrelation chart shows a clear cut off at around 60. Therefore, I conclude that the sales growth rate is an autoregression process with a short-term memory no longer than 60 months (5 years).

Preliminary Conclusion

Periodogram indicates that the sales growth rate follows a periodic pattern, with most prominent periods of 12 months and 3 months. The periodic pattern of 12 months indicates a strong annual seasonality, and the periodic pattern of 3 months indicate a quarterly seasoanlity. Recall that in the model set up by IHS, monthly sales is added together every 3 months, resulting in quarterly sales. The quarterly seasoanlity discovered by the periodogram appears to justify the usage of 3-month moving total.

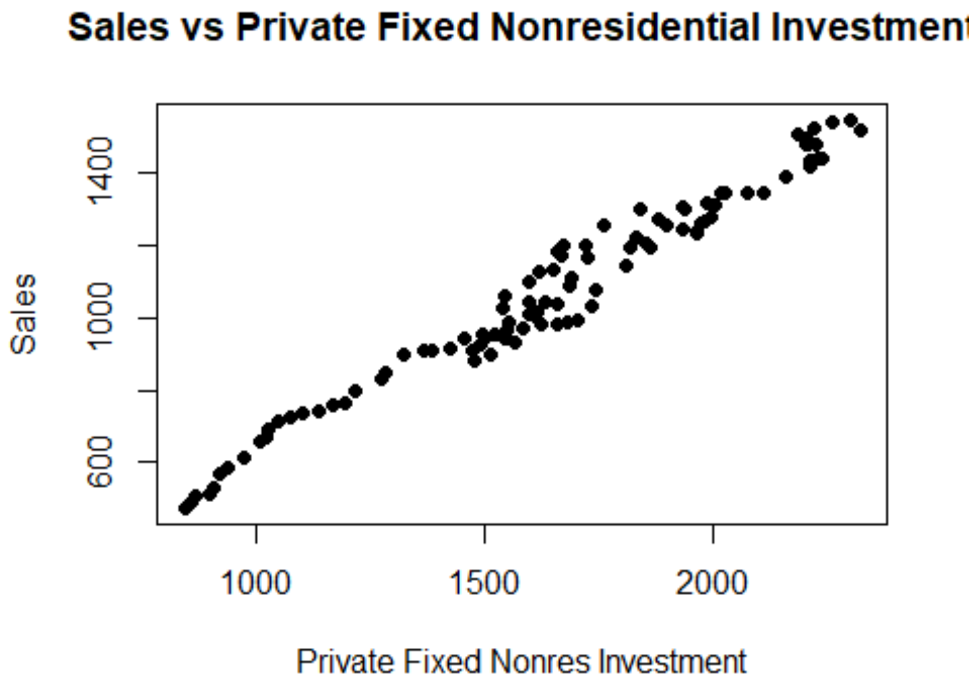
In the next part, I will use 3-month moving total, i.e. quarterly sales, as unit of analysis. I will also take a look at the model set up by IHS, and attempt to validate it.

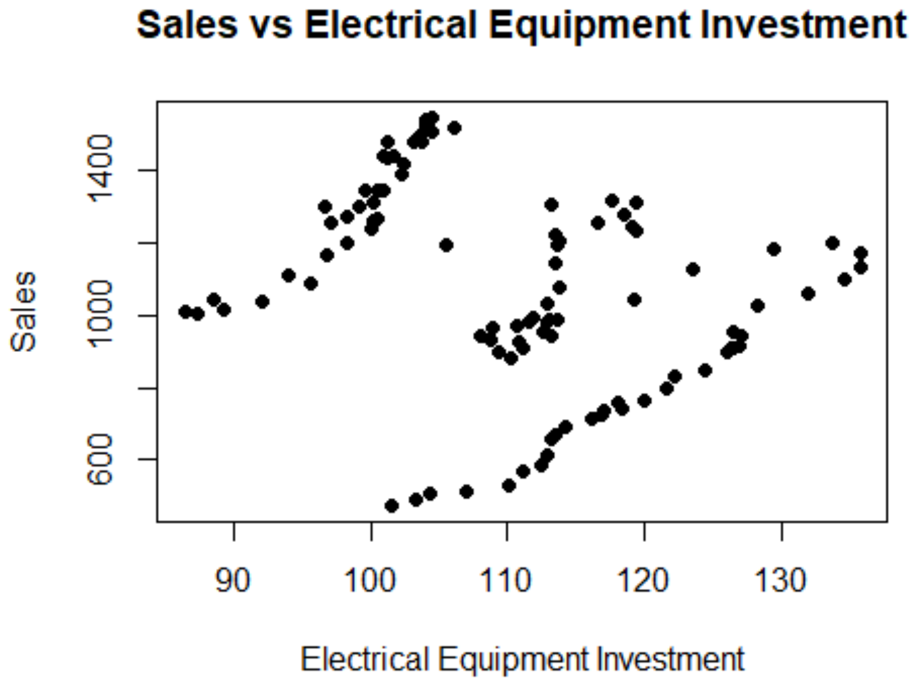
Ordinary Linear Regression

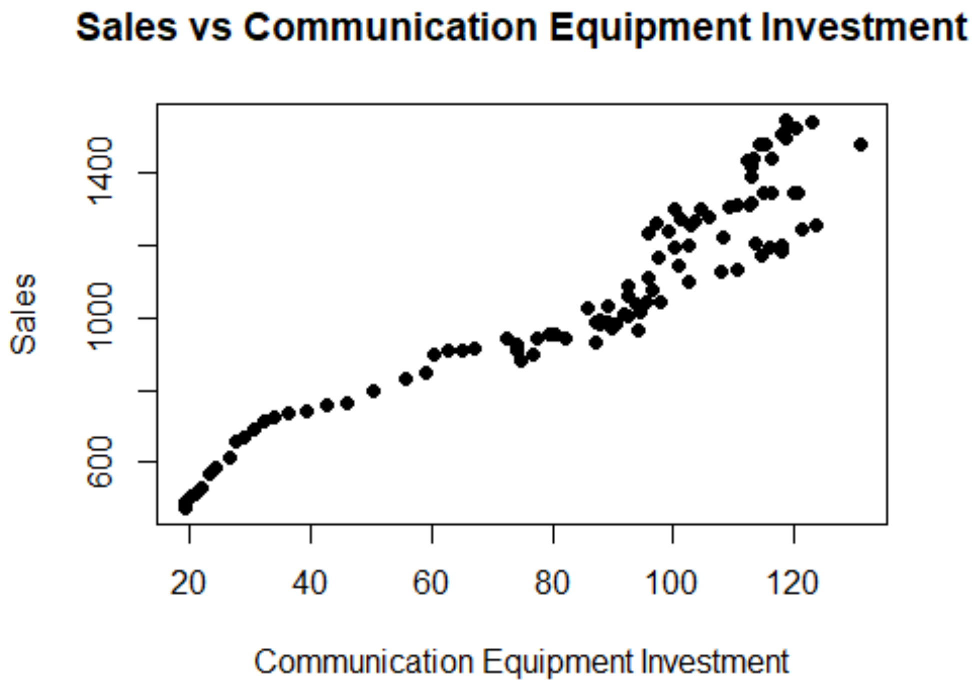
The IHS model is not a purely time-series autoregression model. In addition to sales data, it also employs macroeconomic indicators such as national investment in various sectors, population in employment, etc. All data in IHS's model are quarterly.

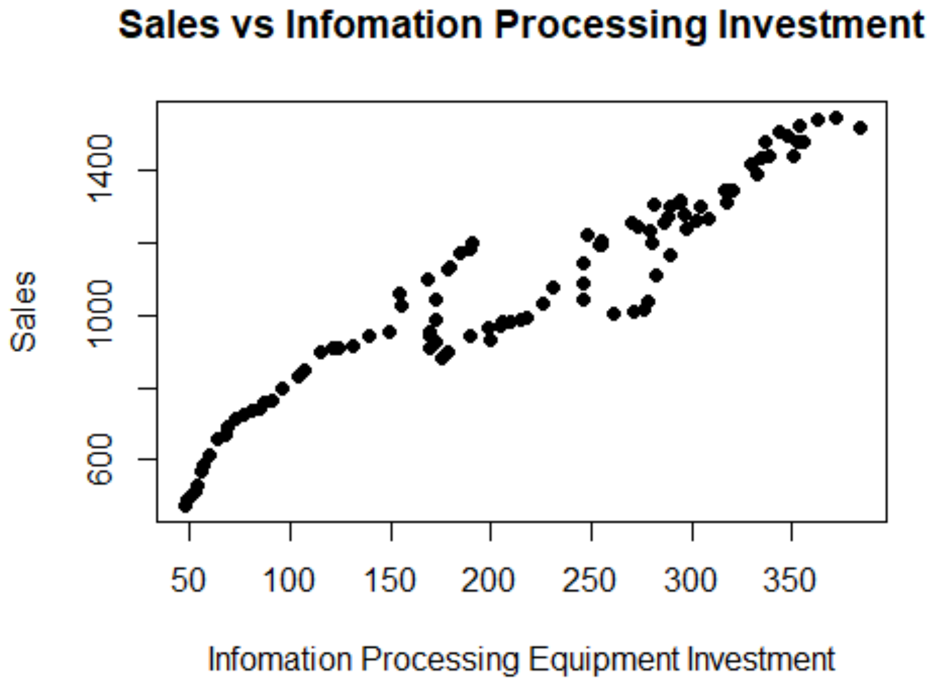
Note that IHS's model's response variable is the first-differenced log sales, i.e. the sales growth rate. I will ignore the transformation for now, and use the sales on its orginal scale as response variable.

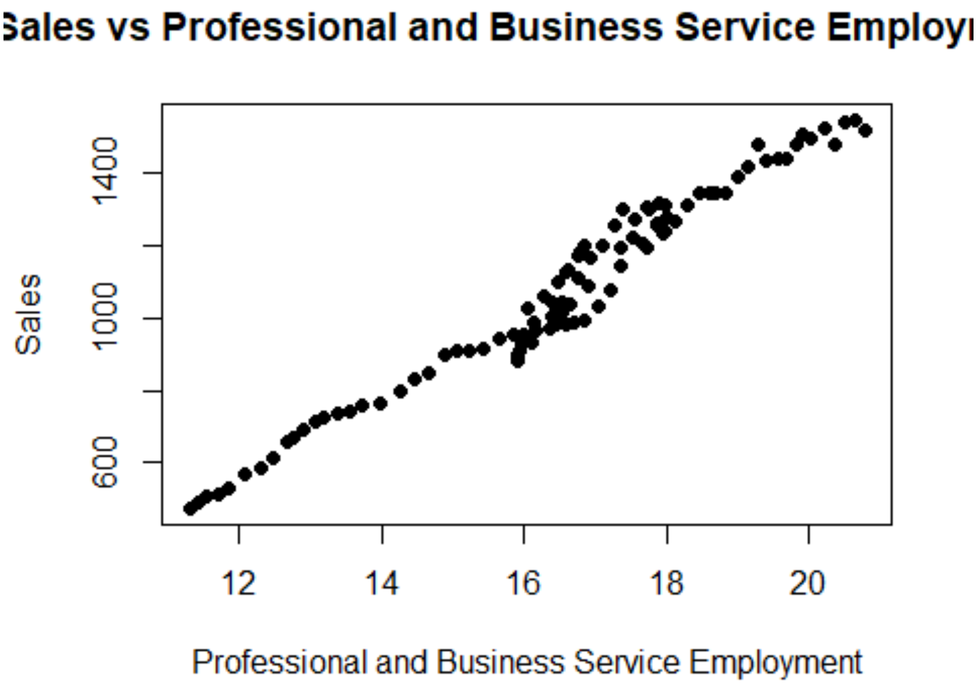
First, let's inspect the relationship between sales and several predictory variables:

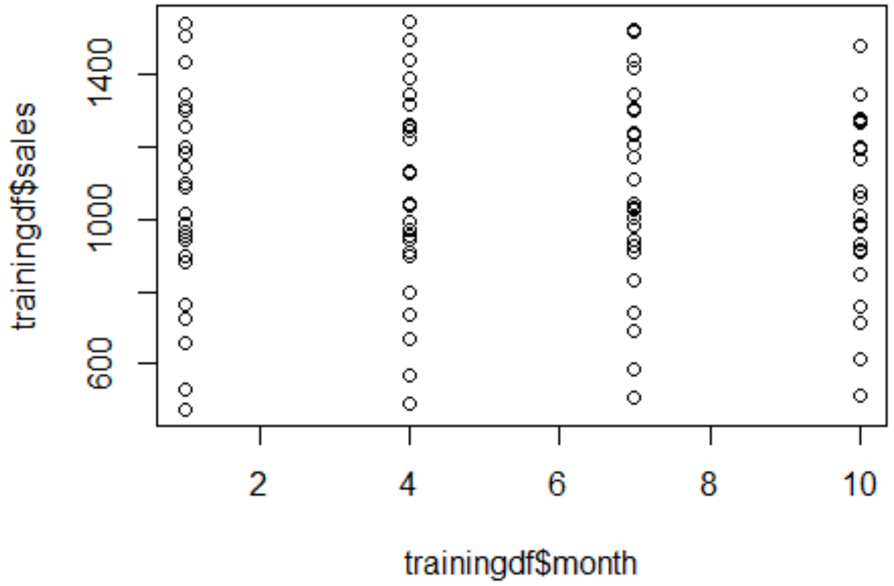

Sales vs Private Fixed Nonresidential Investment
Sales
600
1000
1400
1000
1500
2000
Private Fixed Nonres Investment


Sales vs Electrical Equipment Investment
Sales
600
1000
1400
90
100
110
120
130
Electrical Equipment Investment


Sales vs Communication Equipment Investment
Sales
600
1000
1400
20
40
60
80
100
120
Communication Equipment Investment


Sales vs Infomation Processing Investment
Sales
600
1000
1400
50
100
150
200
250
300
350
Infomation Processing Equipment Investment

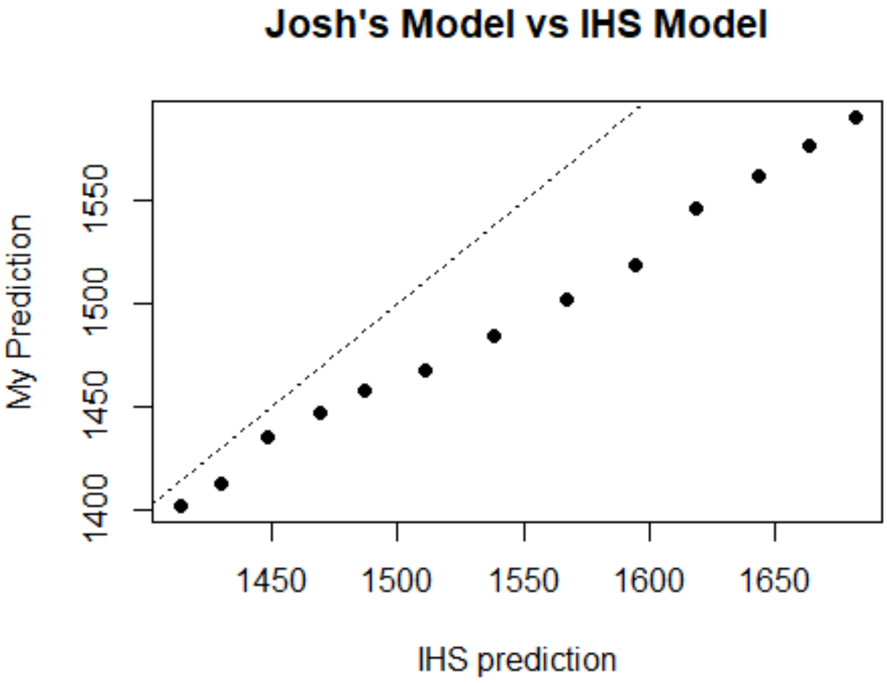




The scatter plot above shows there is a very strong linear relationship between sales and each of the following predictory variables: real private fixed nonresidential investment, Communication Equipment Investment, Infomation Processing Equipment

`Investment`, and `Professional and Business Service Employment`. Now, I will set up a linear model with data from the training dataset.

```
##
## Call:
## lm(formula = I((sales/deflator)) ~ month + rpf + ce + ipe + pbs,
##     data = trainingdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.508 -26.458  -1.076  28.817  74.532
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -405.59020  109.77860  -3.695 0.000372 ***
## month         -0.20924    1.13184  -0.185 0.853735
## rpf            0.25211    0.09392   2.684 0.008605 **
## ce             2.87542    0.38130   7.541 3.05e-11 ***
## ipe           -1.58300    0.16626  -9.521 2.11e-15 ***
## pbs           71.57238   14.61810   4.896 4.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.58 on 93 degrees of freedom
## Multiple R-squared:  0.9696, Adjusted R-squared:  0.968
## F-statistic: 593.4 on 5 and 93 DF,  p-value: < 2.2e-16
```



```
## [1] 3401.775
```

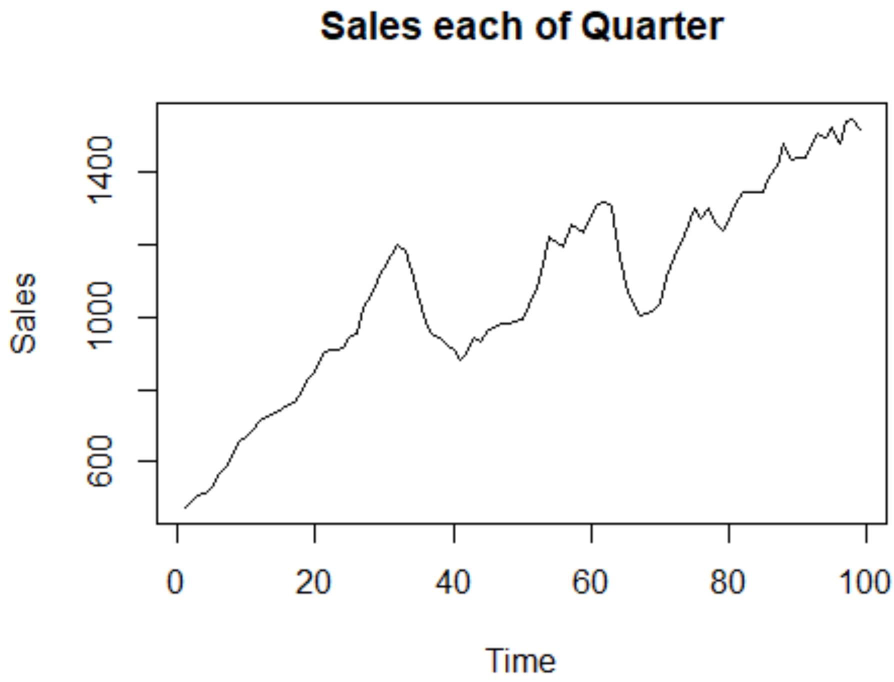
The regression model works very well on the training set. It captures the linear relationship between the response variable and the predictory variables. Predicted values based on the training set match nicely with the real value. But when we make prediction based on the training set, the result is consistently lower than the prediction made by IHS.

Note that the regression model above is not a time series regression: with each observation being a quarter, the residual of each observation is assumed to be uncorrelated. Next, I will apply differencing to data. As a result, residuals of regression will be correlated. Let's see if the new model performs better.

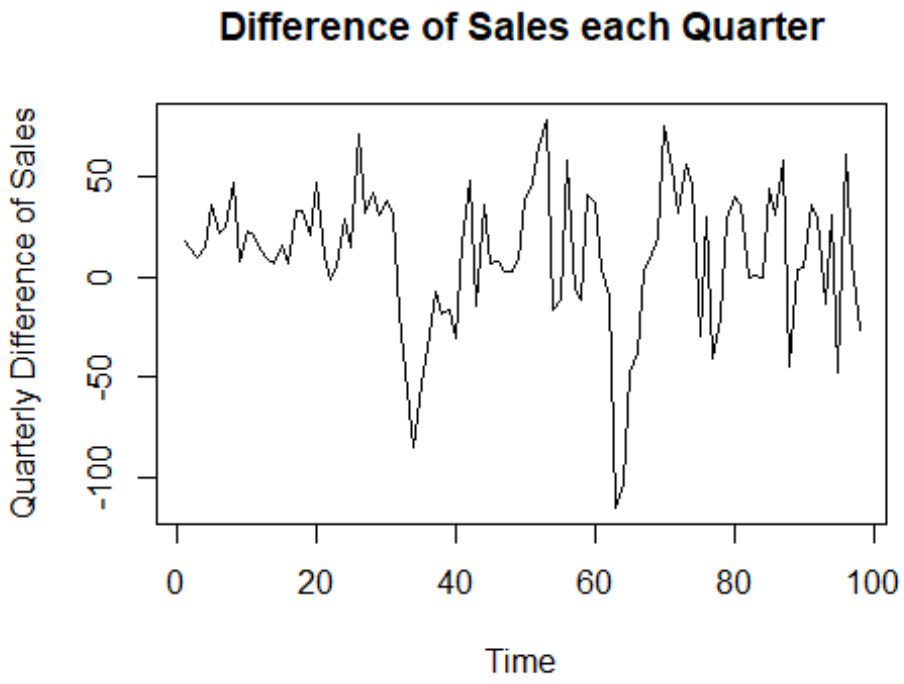
ARIMAX (Autoregression Integrated Moving Average with Exogeneous Predictors)

[NOTE: This part, as well as all following parts, is incomplete.]

To perform time series analysis with ARIMAX model, the response variable is required to be stationary, meaning (1) as random variables, each data point fluctuates around the same center value, and (2) the variance of each data point is the same. With these two criterion in mind, let's take a look at the quarterly sales data, from the first quarter of 1993, to the third quarter of 2017.

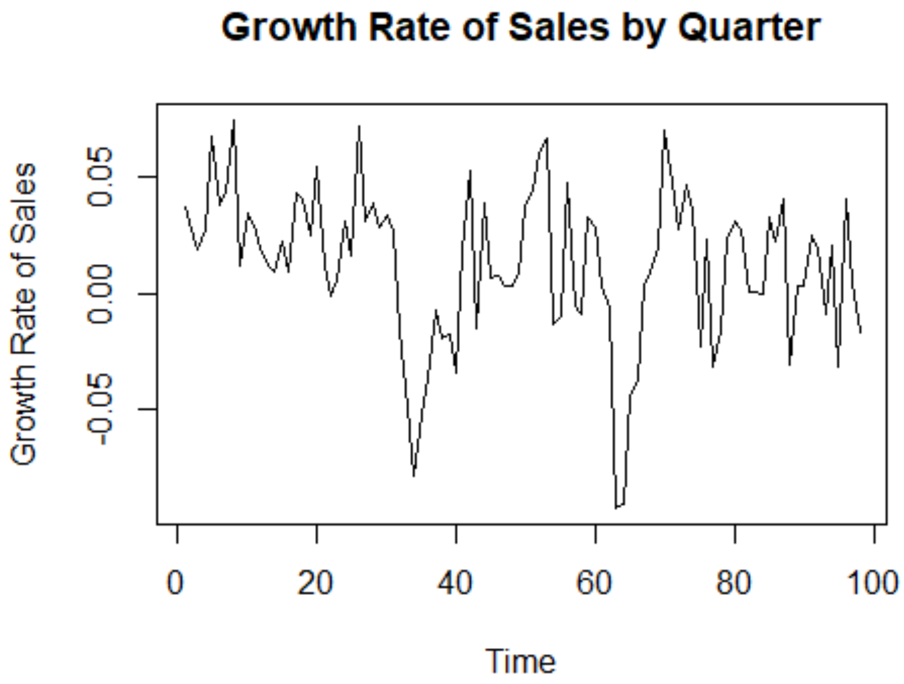


Clearly, quarterly sales has been increasing for the last two decades, and it does not seem to have the same mean. What if we take the first difference of quarterly sales?



The first difference of quarterly sales appears to be more stable than the sales, but I still see the variance grows bigger over time. Now, I will calculate the quarterly sales growth rate, and see if it appears to follow a stationary process. Recall that the growth rate of the T-th period is given by

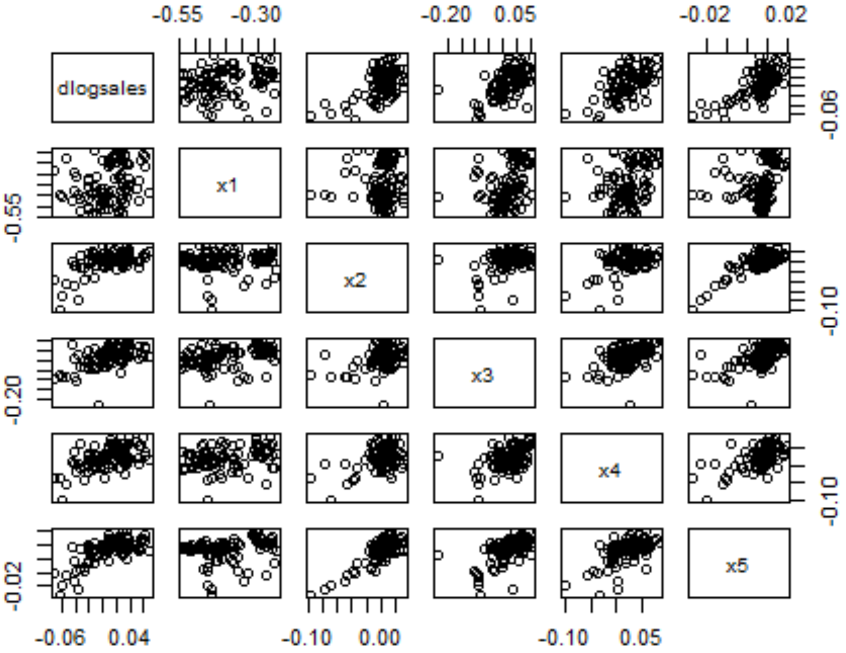
$$r = \ln \frac{sales(T)}{sales(T-1)},$$



TRY ARIMAX next.

```
##   year month  date quarter    sales   salesP      rpf     rpfP       ee
## 1 1993     1 33970  1993Q1 473.0535 473.0535 845.2906 845.3341 101.5192
## 2 1993     4 34060  1993Q2 490.7864 490.7864 861.7915 861.8383 103.2300
## 3 1993     7 34151  1993Q3 504.5035 504.5035 868.8031 868.7988 104.3165
## 4 1993    10 34243  1993Q4 514.1021 514.1021 900.0304 900.0358 106.9911
## 5 1994     1 34335  1994Q1 528.5214 528.5214 910.1730 910.1730 110.1151
## 6 1994     4 34425  1994Q2 564.9978 564.9978 924.1350 924.0821 111.0895
##        eeP       ce      ceP      ipe     ipeP      pbs     pbsP  deflator
## 1 101.5192 19.15413 19.15413 47.12190 47.12190 11.29333 11.29333 0.8273134
## 2 103.2300 19.44207 19.44207 48.12575 48.12575 11.42233 11.42233 0.8275002
## 3 104.3165 20.16643 20.16643 51.19989 51.19989 11.54200 11.54200 0.8287865
## 4 106.9911 21.10430 21.10430 52.13501 52.13501 11.71467 11.71467 0.8303958
## 5 110.1151 21.90053 21.90053 53.81758 53.81758 11.85033 11.85033 0.8349243
## 6 111.0895 23.31690 23.31690 55.18516 55.18516 12.06433 12.06433 0.8402279
##   deflatorP
## 1 0.8273134
## 2 0.8275002
## 3 0.8287865
## 4 0.8303958
## 5 0.8349243
## 6 0.8402279
```

```
##   year month  date quarter    sales   salesP      rpf     rpfP       ee
## 1 2017    10 43009  2017Q4 1552.681 1536.689 2345.279 2341.323 106.4183
## 2 2018     1 43101  2018Q1 1571.494 1560.605 2365.646 2349.320 107.0410
## 3 2018     4 43191  2018Q2 1593.624 1588.130 2368.010 2354.690 107.7389
## 4 2018     7 43282  2018Q3 1619.371 1612.330 2379.332 2365.614 108.4554
## 5 2018    10 43374  2018Q4 1644.638 1636.632 2393.860 2384.012 109.0554
## 6 2019     1 43466  2019Q1 1680.691 1672.766 2415.153 2406.790 109.8194
##        eeP       ce      ceP      ipe     ipeP      pbs     pbsP deflator
## 1 105.1160 118.9284 118.8256 390.3890 380.8267 20.88560 20.87936 1.098341
## 2 105.7605 119.5559 119.4526 393.2560 380.1193 20.97941 21.07875 1.098841
## 3 106.4808 120.2754 120.1714 389.0461 381.5640 21.17294 21.27726 1.100221
## 4 107.2373 120.9786 120.9599 391.1265 384.4120 21.32489 21.39812 1.102255
## 5 107.8925 121.7484 121.7104 393.1832 387.0598 21.43654 21.49063 1.106028
## 6 108.6735 122.5239 122.4955 400.0148 392.1137 21.59330 21.67078 1.112235
##   deflatorP
## 1  1.096609
## 2  1.097107
## 3  1.098484
## 4  1.100514
## 5  1.104284
## 6  1.110486
```



```
##
## Call:
## lm(formula = dlogsales ~ x1 + x2 + x3 + x4 + x5)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060949 -0.011391 -0.000925  0.011876  0.058886
##
```

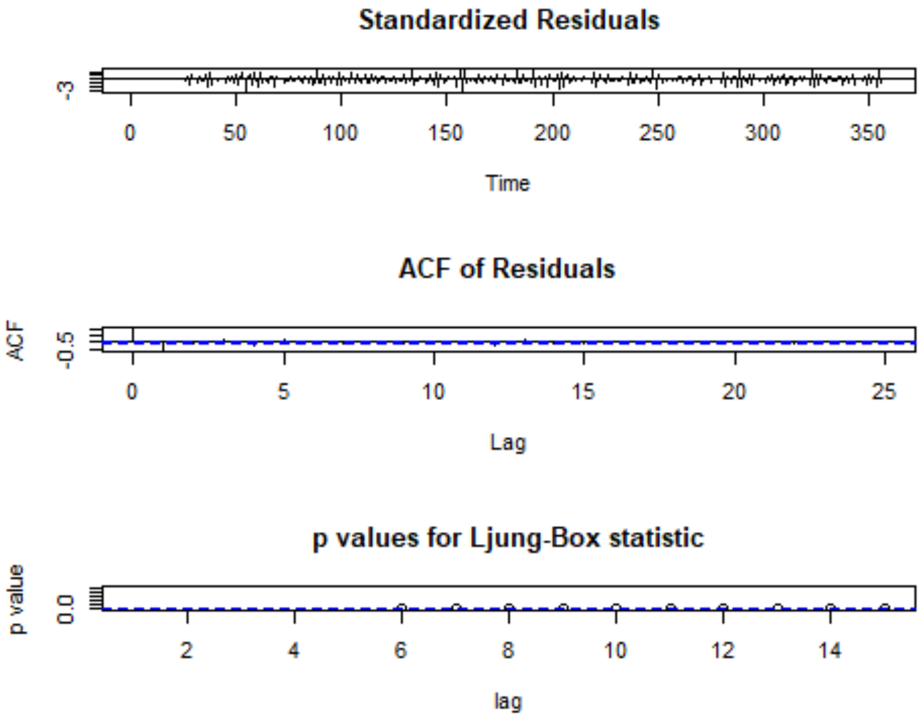
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004145   0.012058  -0.344  0.73184
## x1           0.004803   0.026732   0.180  0.85780
## x2           0.053185   0.177636   0.299  0.76531
## x3           0.119840   0.049657   2.413  0.01779 *
## x4           0.151183   0.086935   1.739  0.08538 .
## x5           1.542700   0.521849   2.956  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02048 on 92 degrees of freedom
## Multiple R-squared:  0.5491, Adjusted R-squared:  0.5246
## F-statistic:  22.4 on 5 and 92 DF,  p-value: 1.252e-14
```

Next, I will apply regression to the historical sales growth. First, let's test if the average of sales growth is zero.

```
## [1] "p-value between 5% and 10%"
```

Test result shows that we can reject the zero-mean hypothesis at the 0.1 significance level. For further analysis, I will remove the average from the sales growth rate.

Recall that autocorrelation chart shows a gradually tailing off pattern, and the partial-autocorrelation chart shows a cut-off at a lag around 60. Considering the periodogram shows a prominent periodicity at 12 months, I intend to try the multiplicative seasonal autoregression integrated moving average model, denoted by $ARIMA(p, d, q) \times (P, D, Q)_s$, where



Standardized Residuals

ACF of Residuals

p values for Ljung-Box statistic