**Morgan Blein**

**Forecasting housing prices in San Francisco**

**Objective:** building an objective forecasting model for the San Francisco SP-Case-Shiller home prices index using publicly available data (see part III for more details about these variables).

**Tools**: STATA/ excel

Stata do file: a script file containing all the information to run the desired analysis in Stata.

[Here is a link to the Do file for this project.](https://www.dropbox.com/s/imxw7334l9144fu/BLEIN_forecasting_home_prices.do?dl=0)

[Here is a link to the data compiled for this project](https://www.dropbox.com/s/77tqa3mz9oxhr16/BLEIN_forecasting_home_prices.dta?dl=0).

**Method**: Regression

**OUTLINE:**

1. **Introduction**
2. **Seasonality**
3. **explanatory variables**
4. **Causal regressions:**
5. **forecasting model**
6. **Introduction**

Standard and Poor’s is rating agency, mostly famed by the stock trading index S&P 500. However, stock markets is not the only thing they monitor. Starting in the 1970s, they developed a housing prices index, for major metropolitan areas. Today, they generate these indexes for the 20 largest cities in the United States. Today, we are going to focus on the San Francisco home prices index. This index starts in January of 1980.

Here are facts about our data:

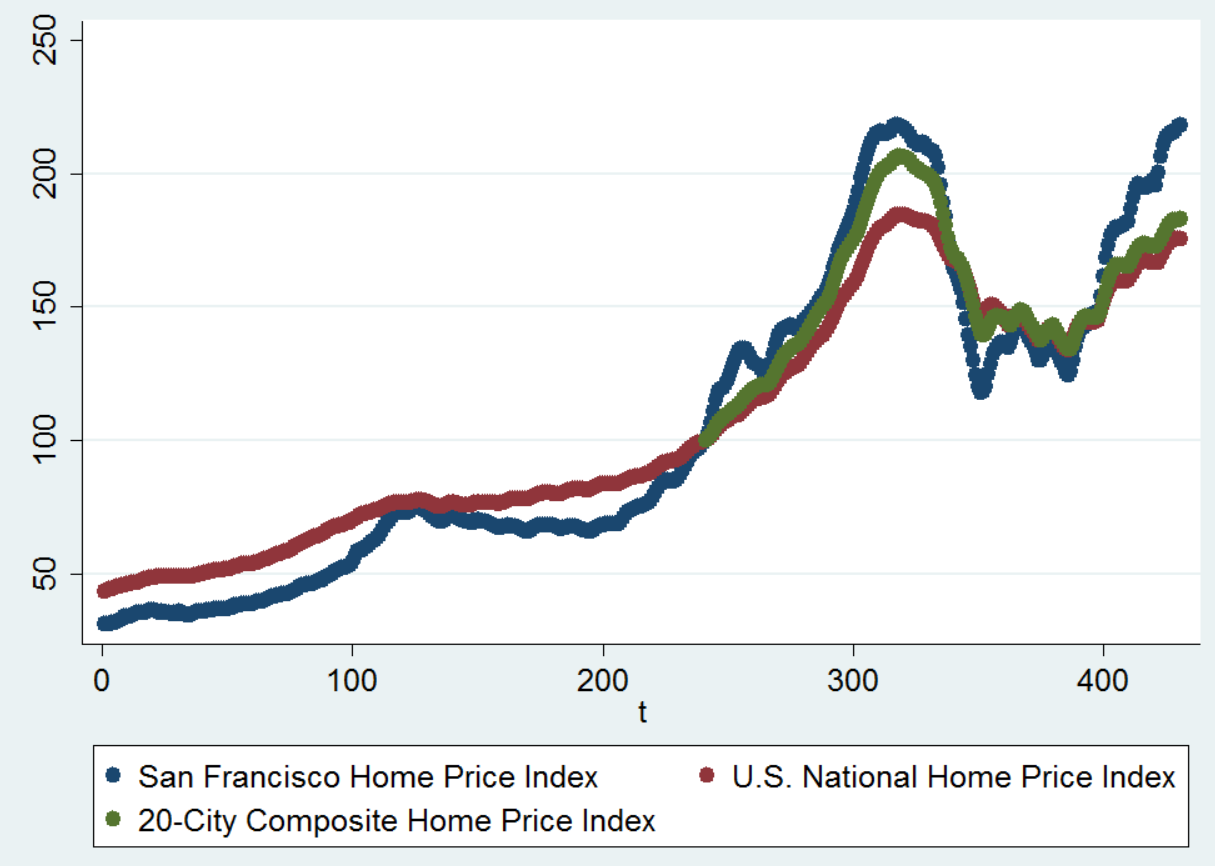
* Number of observation: 431.
* Each observation corresponds to a value for a given month.
* Time period range: from January 1980, until November 2015.
* Index Jan 2000=100

The main area of focus of this study will be first seasonality. Do any seasonal pattern exist? Do they have an impact on the home index values? Then, we will choose several external variables in order to run a causal regression on the SF home price index. In other words, what are the driving forces behind the value of this index? In order to do so, we are going to carefully choose variables and discuss their consistency and unbiaisness. The final step of our analysis will consist of a forecasting model, in order to generate potential values for February 2016 and October 2016.

1. **Seasonality:**

A section of seasonality that includes a regression table, any F-statistic calculation, and a written narrative description of the seasonality patterns

1. Plot home price index/ t (period unit in month)



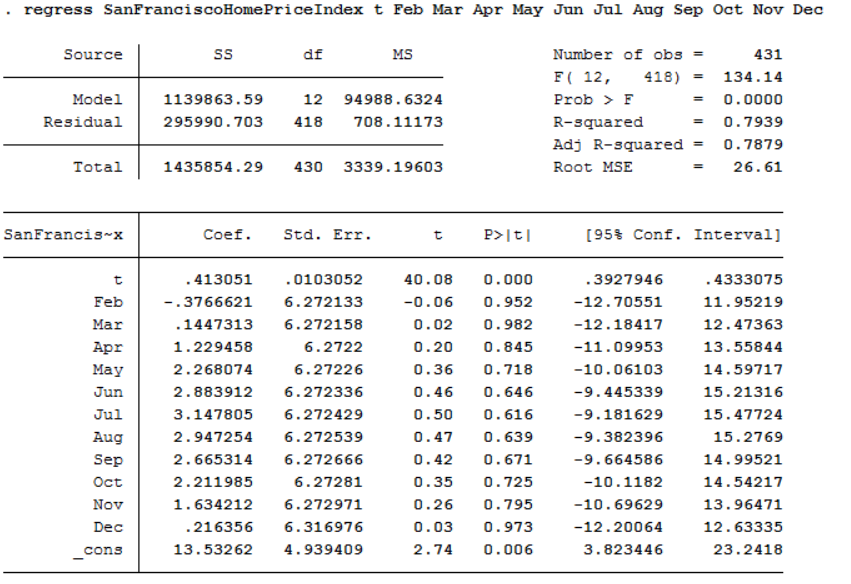
We can see from the graph that the values of all 3 indexes (San Francisco home index, U.S National home index, 20-city composite home index) all follow the same trend. There are obvious variations, but as a whole, the curves tend to behave in the same way.

Over time, the indexes for home prices grows. This intuitively makes sense: it is common knowledge that a house today is worth a lot more than in the 1980. I suspect the value of t (time unit in month over the whole period) to be heavily correlated to the San Francisco home price index (and other indexes). The relationship seems pretty linear in nature (or maybe exponential, more calculation will be done in the next section) especially up to period T equals to about 300. After, this we can notice a severe decrease in the price index. That period corresponds to May 2007 or the US housing bubble burst. This trend drove prices down until March 2012, when index values start rising again.

1. **Linear seasonality trend:**

yt = α0 + α1t + et , t = 1, 2, etc..

We can infer from the graph the influence of time over the price indexes. However, we still do not know if any other seasonality during the year from month to month occurs. I order to confirm those assumption and get answer, we regress the variable **SanFranciscoHomePriceIndex** on a linear time trend and 11 monthly dummy variables, using January as the base month.



* Number of observations: 431
* Adjusted R-squared: 0.79
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

The regression equation sums up to:

* SanFranciscoHomePriceIndex = 13.53 + .413 t – 0.376 feb - .1447 mar + 1.22 apr + 2.26 may + 2.88 jun + 3.14 jul + 2.95 aug + 2.66 sep + 2.21 oct + 1.63 nov + 0.21 dec

It feel when looking at the coefficients that months would have a great influence on the San Francisco home index price. However, I think these variables are not statistically significant. Their P values are very much over 5% ranging from 0.616 to 0.982). Their “t” values are very low and the standard error much too high.

T however has a P-value of 0. We can assume the high R squared comes from this time over period’s trend (since all other variables are non-significant)

Key findings: seasonality over t exists, however no seasonal trend for months seem to exist.

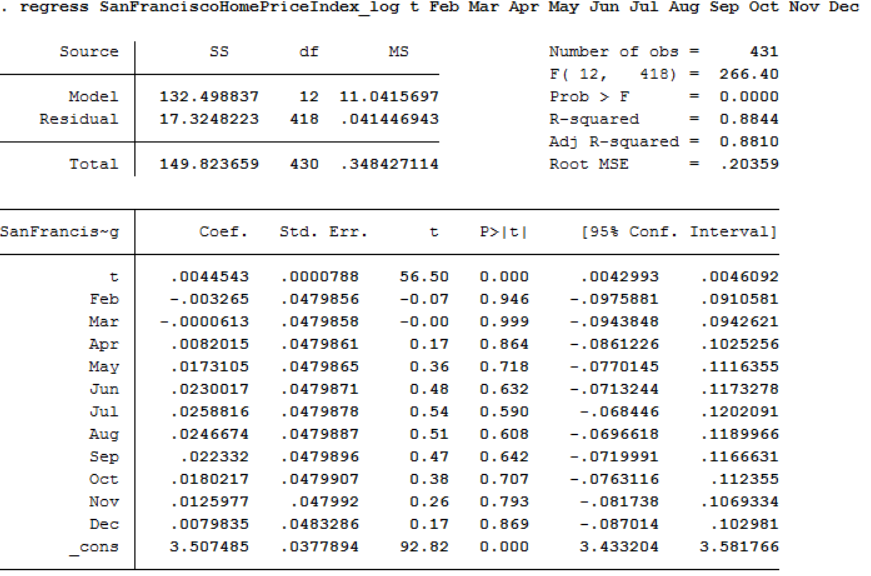
1. **Exponential trend**

log(yt) = α0 + α1t + et ,t = 1, 2

To decide whether to follow a linear or exponential trend model, we need to create a log variable based on SanFranciscoHomePriceIndex. We call this variable SanFranciscoHomePriceIndex\_log and it follows this equation:

SanFranciscoHomePriceIndex\_log = log(SanFranciscoHomePriceIndex)

We now run a regression trend model based on this new variable:



* Number of observations: 431
* Adjusted R-squared: 0.88
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

The R squared went up from 79% to 88%. All variables other than t seem statically irrelevant, same as before. However this seems to give a better base for the analysis.

This will have to be kept in mind when doing the causal analysis. There are 2 approaches I took to deal with this time trend:

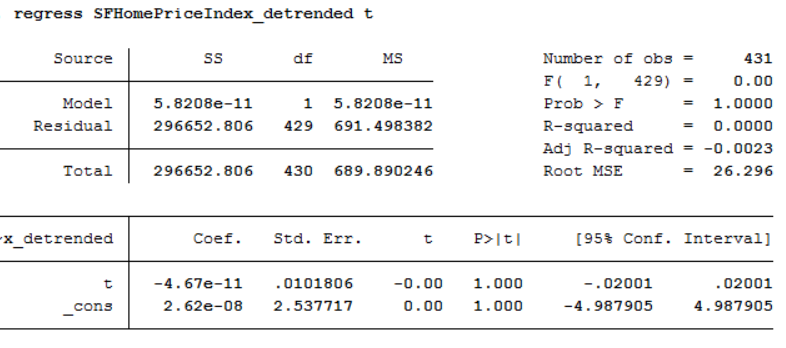
Method 1:

* Keep the t variable in the regression model, use the variables SanFranciscoHomePriceIndex or SanFranciscoHomePriceIndex\_log to run the regression

Method 2:

* Use STATA tools to detrend SanFranciscoHomePriceIndex or SanFranciscoHomePriceIndex\_log. New variables named SFHomePriceIndex \_detrended and SFHomePriceIndex\_log\_detrended will be created. Alos use stata to detrend the independent variables.

We can make sure these variables are de-trended by running a new regression on one of them:

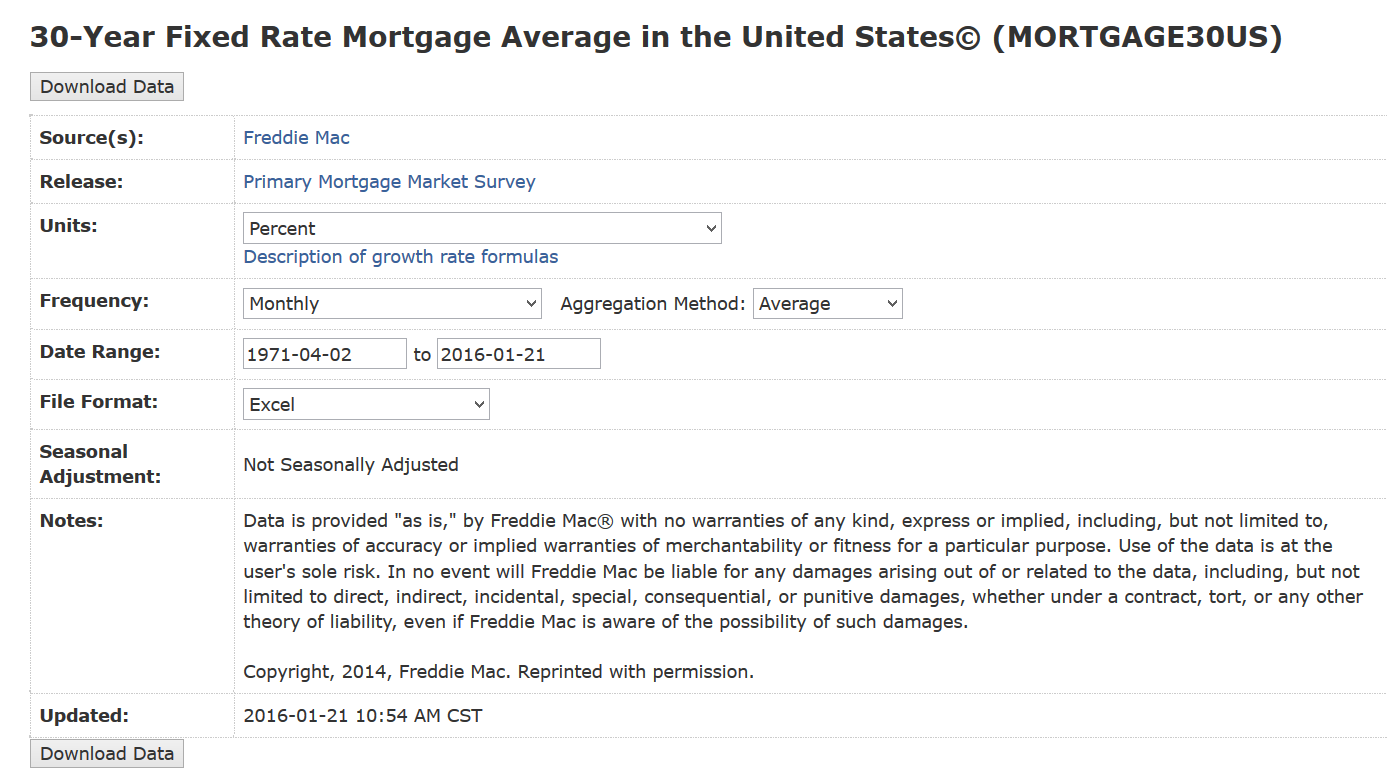


The model has no statistical significance. Time therefore has no influence on the new variable anymore.

1. **explanatory variables**

In this section, I will list the explanatory variables I included, their sources, as well as providing a justification for their use.

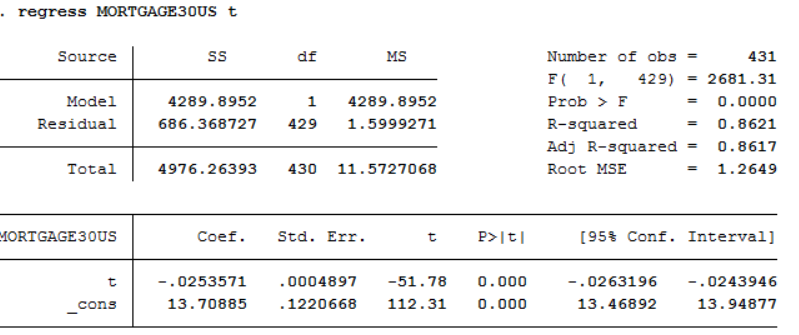
1. [**30**](https://research.stlouisfed.org/fred2/search?st=Fixed+Rate+Mortgage) **years Mortgage rates**



Freddie Mac, 30-Year Fixed Rate Mortgage Average in the United States© [MORTGAGE30US], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/MORTGAGE30US, January 26, 2016.

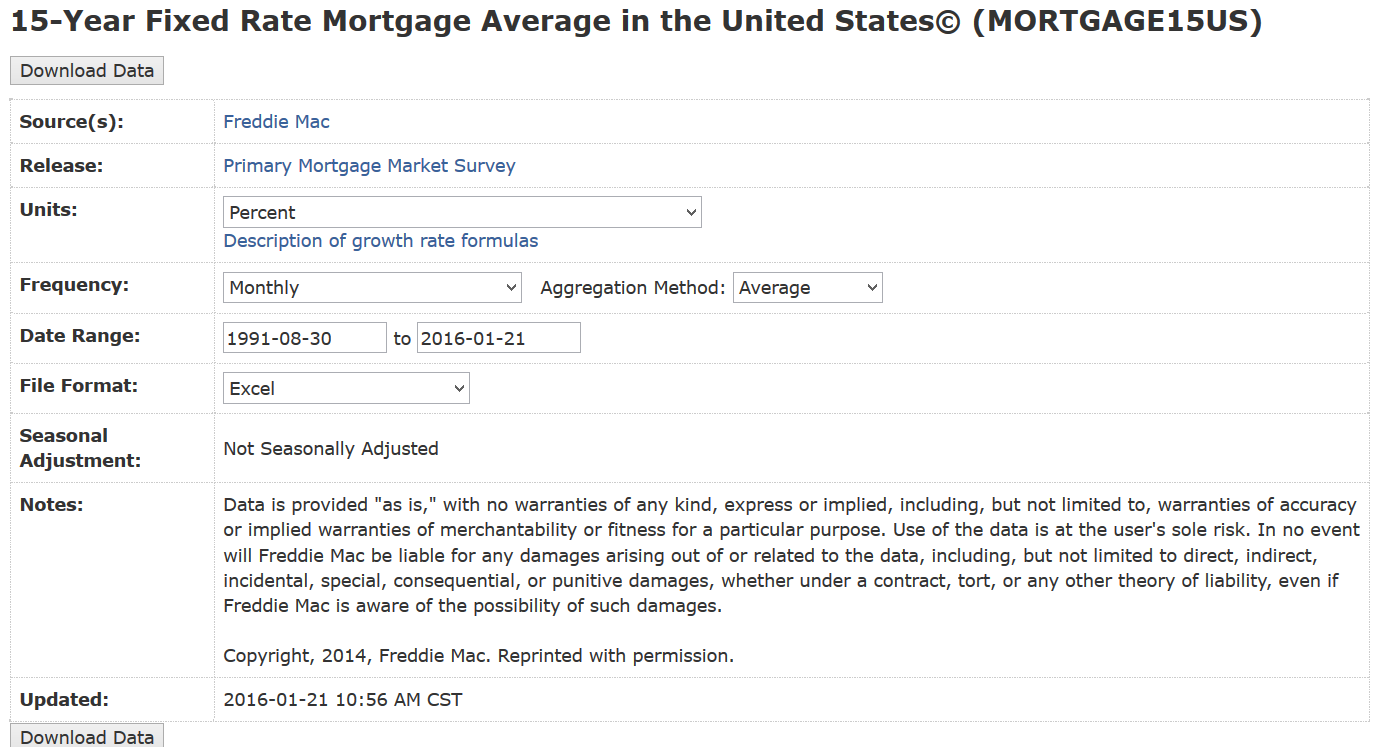
Mortgage rates could have an impact on house prices index, as they are directly to take into account. Most people resort to mortgages when buying a house. The 30 year mortgage is the most popular. Moreover the data given covers the whole period I am interested in. As seen in the screen shot above, the data is sourced monthly (using the average aggregation method) and covers our whole time period range (January 1980 to November 2015)

The data is not seasonally adjusted. When regressing with t (time unit) we get:



The R squared is 86%. We need to detrend this variable.

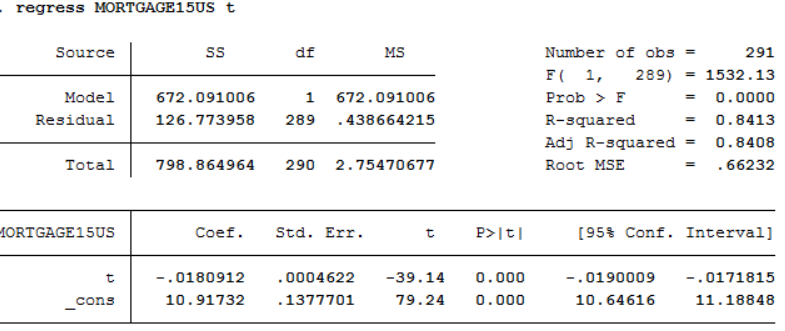
1. **15 years Mortgage rates:**



Freddie Mac, 15-Year Fixed Rate Mortgage Average in the United States© [MORTGAGE15US], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/MORTGAGE15US, January 25, 2016.

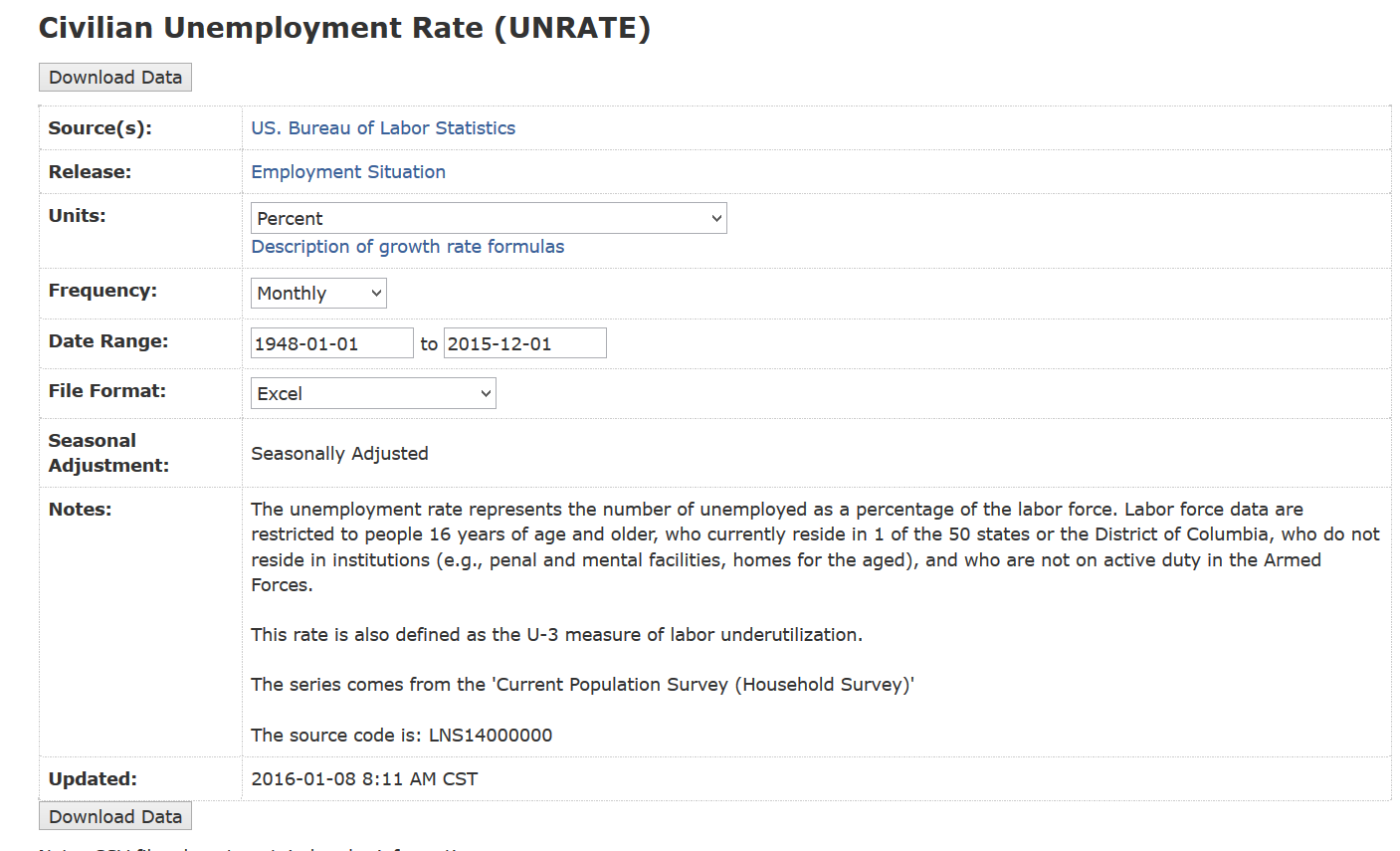
Same explanation as for the 30 year mortgage. I thought it was worth including as some people may choose this option instead of the 30 year mortgage.

The data is not seasonally adjusted. When regressing with t (time unit) we get:



The R squared is 84%. Therefore we need to create a detrended series of values for this variable.

1. **Unemployment rates:**

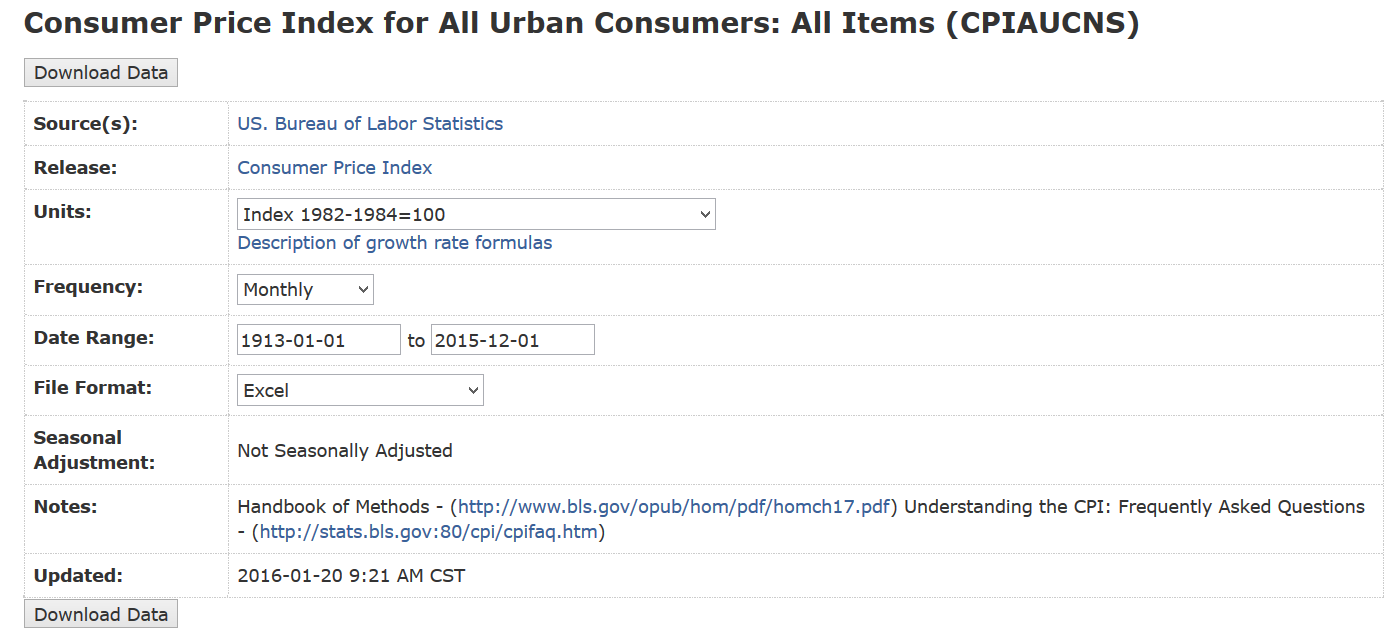


US. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/UNRATE, January 27, 2016.

This is needed as it can also have an impact on home prices. It is a reflation on the general economic wellbeing of a state or area. As mentioned in Moody’s case-shiller-methodology:“The unemployment rate is relevant since the buyers of lower-cost homes tend to be lower income and are thus more sensitive to the local business cycle and job prospects than higher-income households.”

The data is already seasonally adjusted. We do not need to detrend this variable further.

1. **Consumer price index:**

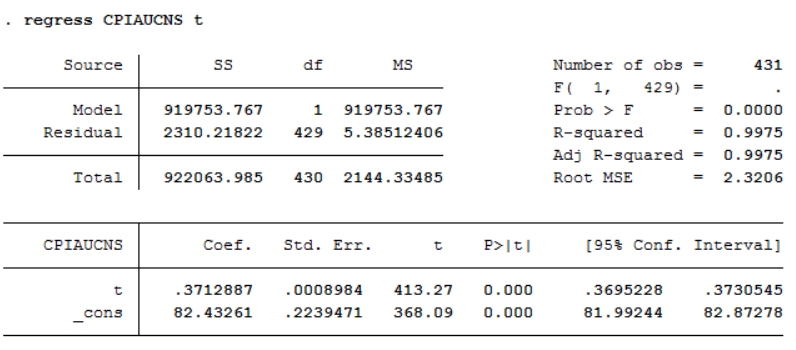


US. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/CPIAUCSL, January 27, 2016.

The Consumer Price Index (CPI) is a defined as: “measure of the average change over time in the prices of consumer items—goods and services that people buy for day-to-day living.” (<http://www.bls.gov/opub/hom/pdf/homch17.pdf>)

It gives a good estimate of how daily products varies in price over time. It will be interesting to compare against as purchase such as a home.

The data is not seasonally adjusted. When regressing with t (time unit) we get:



The R squared is 99%. Therefore we will also create a detrended version of this variable.

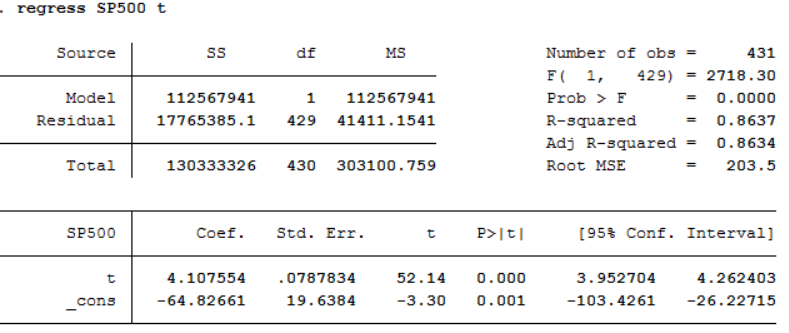
1. **SP500 stock index:**

<https://finance.yahoo.com/q/hp?s=%5EGSPC&a=00&b=3&c=1950&d=00&e=26&f=2016&g=m>

**Note**: the opening price was the index selected.

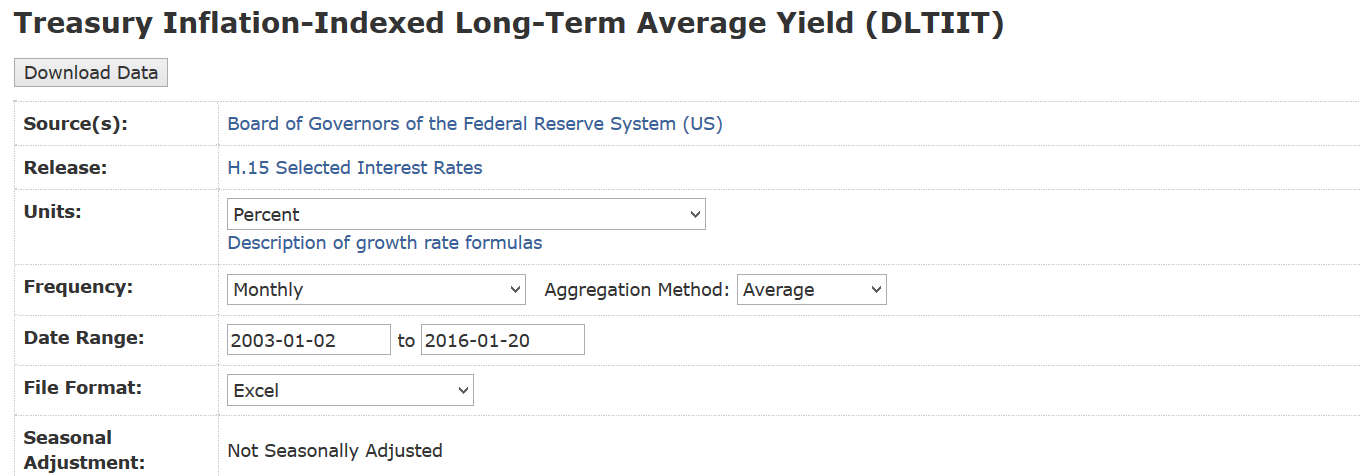
According to: <http://us.spindices.com/>, The S&P 500 is: “widely regarded as the best single gauge of large-cap U.S. equities[…] The index includes 500 leading companies and captures approximately 80% coverage of available market capitalization.” This indicator covers a very wide range to address stock trading overall health and trends. In the United States, it can severely impact the overall economic standing of the nation. Trying to find a correlation with home prices will tell us more about the relationship between the 2 indexes.

The data is not seasonally adjusted. When regressing with t (time unit) we get:



The R squared is 86%. Therefore we will also create a detrended version of this variable.

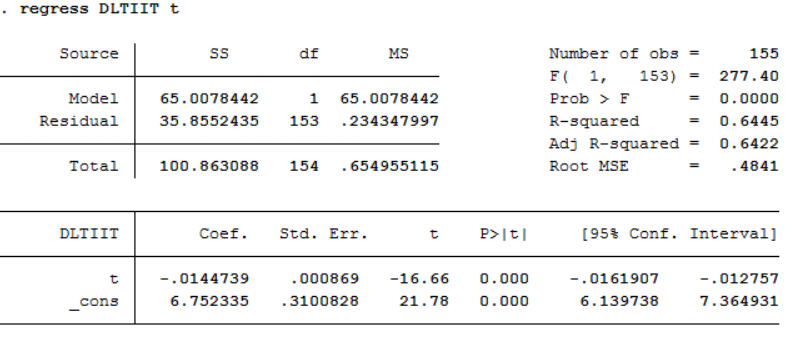
1. **Treasury Average Yield:**



Board of Governors of the Federal Reserve System (US), Treasury Inflation-Indexed Long-Term Average Yield [DLTIIT], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/DLTIIT, January 27, 2016.

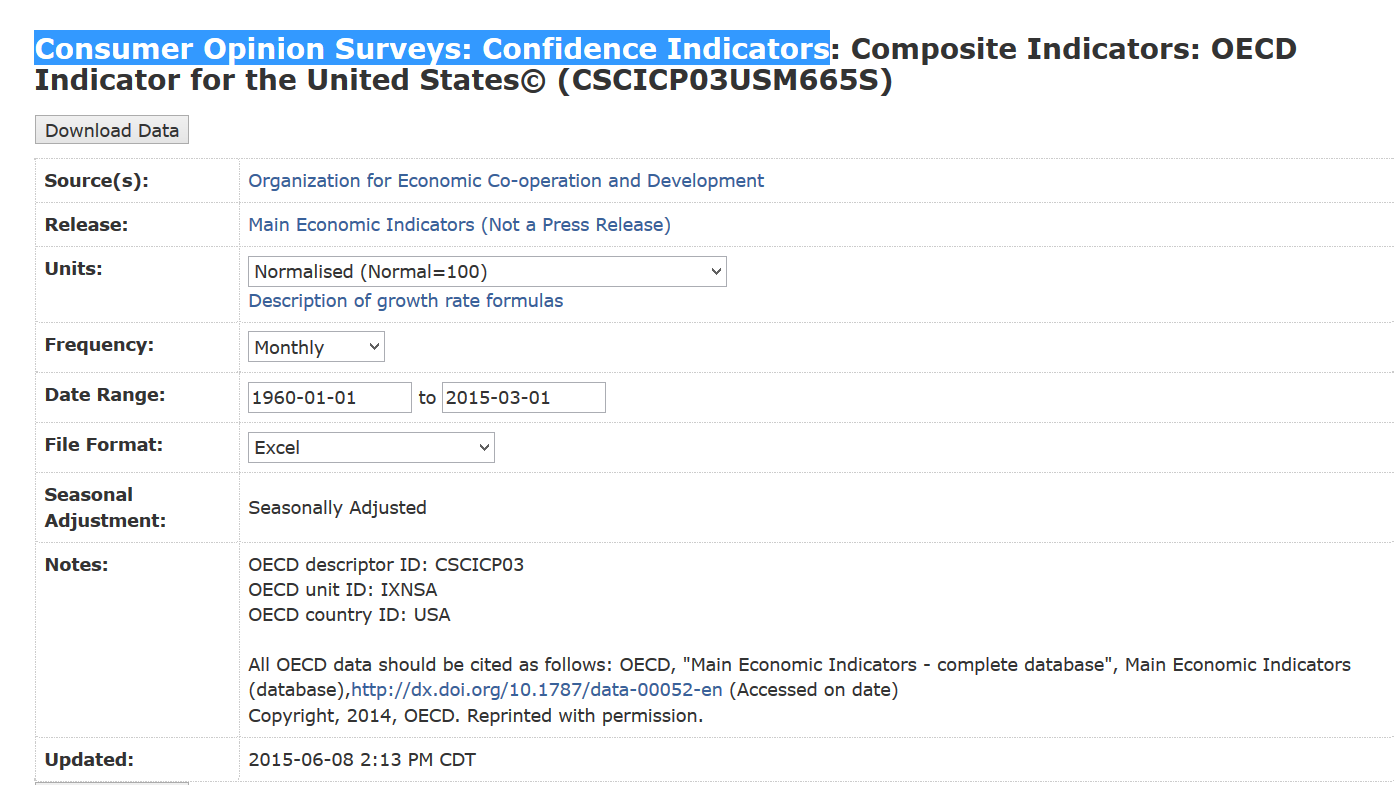
Treasury yield is the interest rate the U.S. government pays to borrow money for different lengths of time. It is considered one of the safest investment due to the stability of the US government. Usually, their yield is very low. The influence of their rates however extends to more than just the US government: “they also influence the interest rates individuals and businesses pay to borrow money to buy real estate, vehicles and equipment.” (<http://www.investopedia.com/terms/t/treasury-yield.asp>)

The data is not seasonally adjusted. When regressing with t (time unit) we get:



The R squared is 64%. This is still high enough to justify detrending the values of this series.

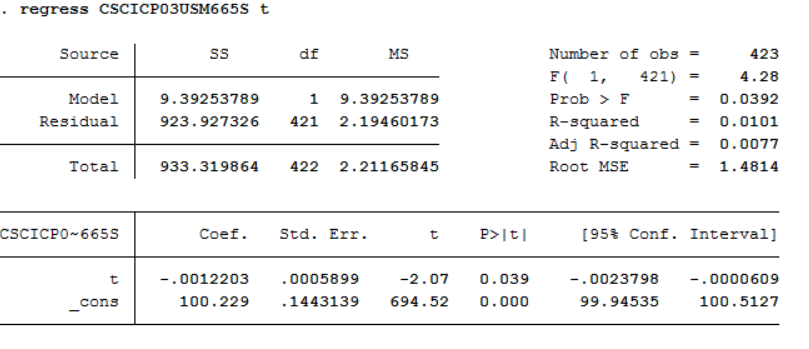
1. **Consumer opinion confidence indicators**



Organization for Economic Co-operation and Development, Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: OECD Indicator for the United States© [CSCICP03USM665S], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/CSCICP03USM665S, January 27, 2016.

This index is very interesting. It is a behavioral index, relying on the psychology of purchase. It shows to confidence level of consumers to make a purchase in the United States. The format is monthly as usual, and the data ranges from the 1960s to early 2015.

The data is seasonally adjusted. When regressing with t (time unit) we get:



The R squared is around 1%. Therefore, no need to create a detrended value. We can see the website did a good job detrending time out of this variable.

1. **SP-Case-Shiller home prices indexes:**

S&P created home price indexes for the 20 largest metropolitan areas in the United States. We collected data for each of these cities (Atlanta, Boston, Chicago, Cleveland, Dallas, Denver, Detroit, Las Vegas, Los Angeles, Miami, Minneapolis, New York phoenix Portland, San Diego, san Francisco, Seattle, Tampa, Washington DC). We also collected the National US average home price index, as well as a composite index for all top 20 cities.

<http://us.spindices.com/indices/real-estate/sp-case-shiller-ca-san-francisco-home-price-index>

This concludes our variables selection section. In short:

* 22 variables for home price indexes
* 7 external econometric variables

1. **Causal regressions:**
2. **First set of regressions**

Section describing your causal regressions, describing its results in both a table and written narrative, and discussing its consistency and unbiasedness

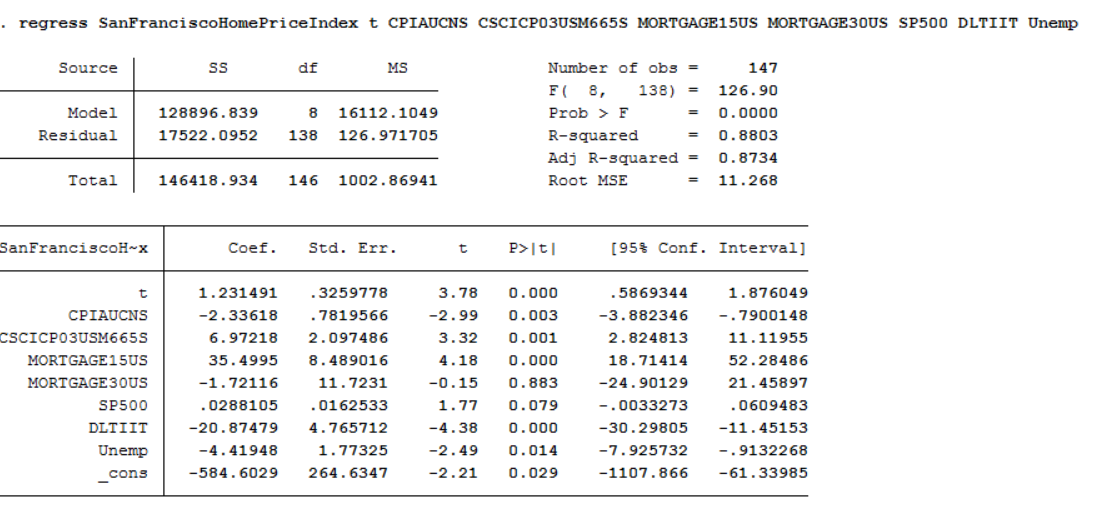
We will run for sets of regression. All explanatory variables (described in the previous section) are the same for these regression, except for the time period (t) that will be added as an explanatory variable for the non-de-trended Y-variables (see end of section II) Seasonality)

Method 1:

1. Using Y= SanFranciscoHomePriceIndex, t included, X variables as they are.
2. Using Y= SanFranciscoHomePriceIndex\_log, t included, X variables as they are.

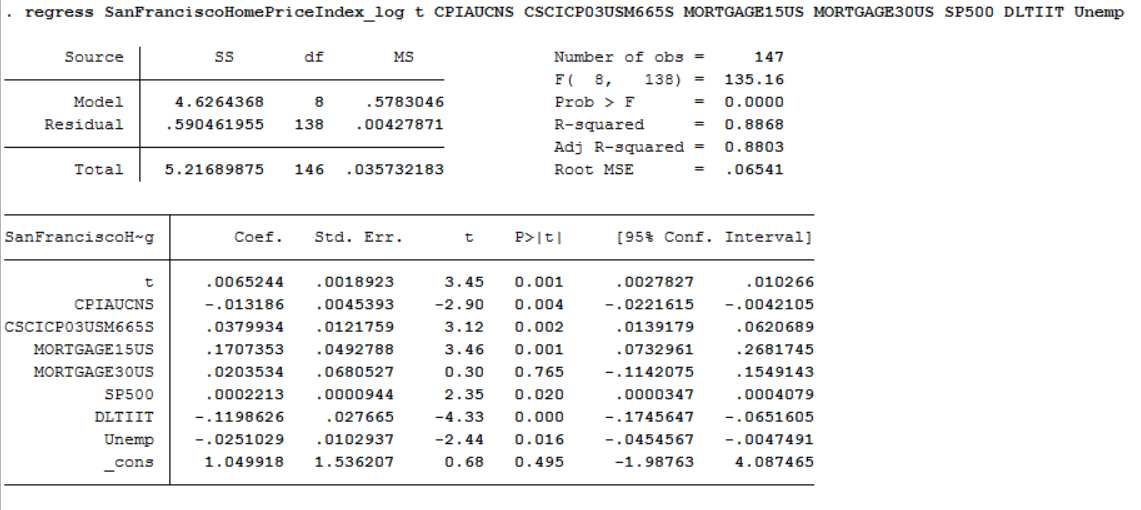
Method 2:

1. Using Y= SFHomePriceIndex\_detrended, X variables detrended
2. Using Y= SFHomePriceIndex\_log\_detrended, X variables detrended
3. Using Y= SanFranciscoHomePriceIndex, t included



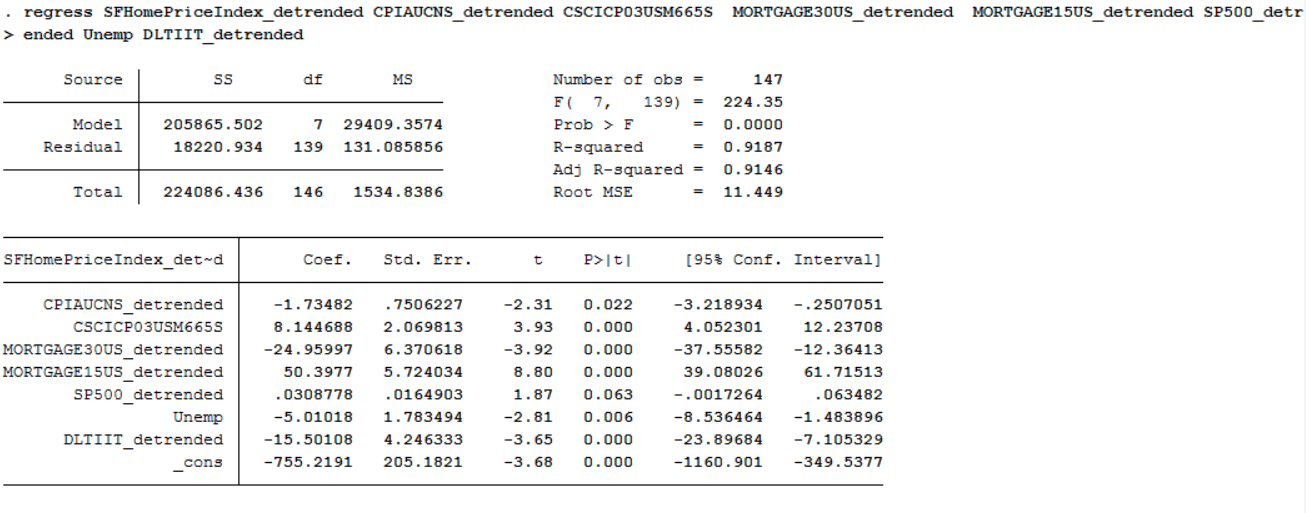
* Adjusted R-squared: 0.88
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

1. Using Y= SanFranciscoHomePriceIndex\_log, t included



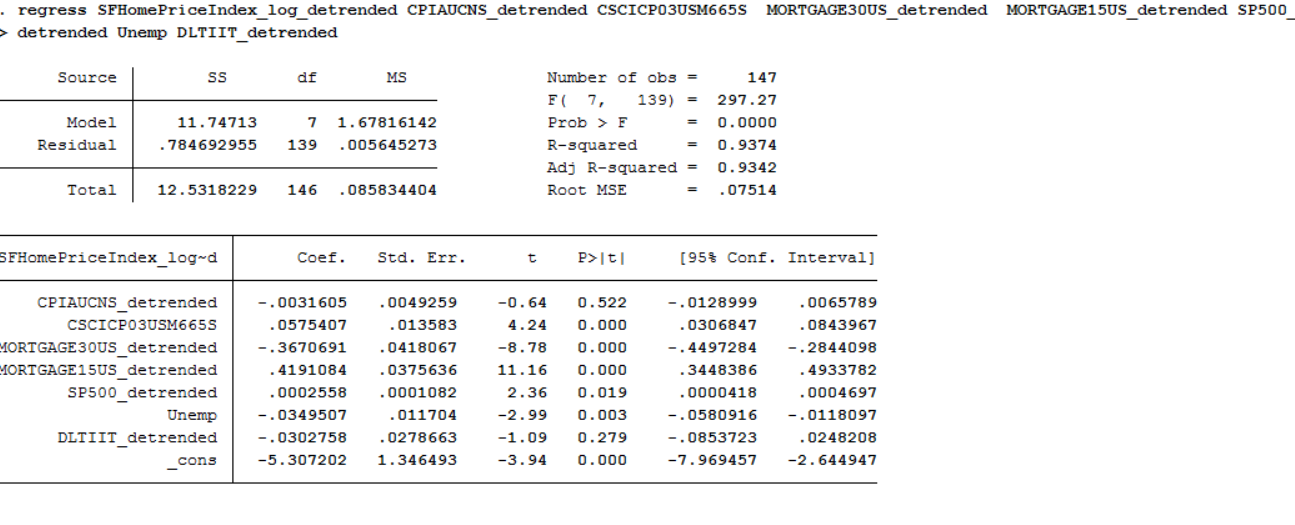
* Adjusted R-squared: 0.88
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

1. Using Y= SFHomePriceIndex\_detrended



* Adjusted R-squared: 0.918
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

1. Using Y= SFHomePriceIndex\_log\_detrended



* Adjusted R-squared: 0.937
* Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

The R squared of each model from a-d is very good: ranging from 88% to 93%. All the models are statistically sound. They best way to choose one here is to look at the endogenous variables. As mentioned in the data collection and sources, two of my variables came seasonally adjusted. They are:

* Consumer opinion confidence indicators (CSCICP03USM665S)
* Unemployment rate.

In my opinion, this creates an issue. In model a) and b), the inclusion of t should take care of the seasonality. Since those are already detrended from source, we are facing an issue. Therefore for consistency, I want to focus on models c) and d).

Out of these 2, I want to take a closer look at model c). In model d), two of my variables may not be statistically significant:

* CPIAUCNS\_detrended, P-value of 0.522
* DLTIIT\_detrended, P-value of 0.279

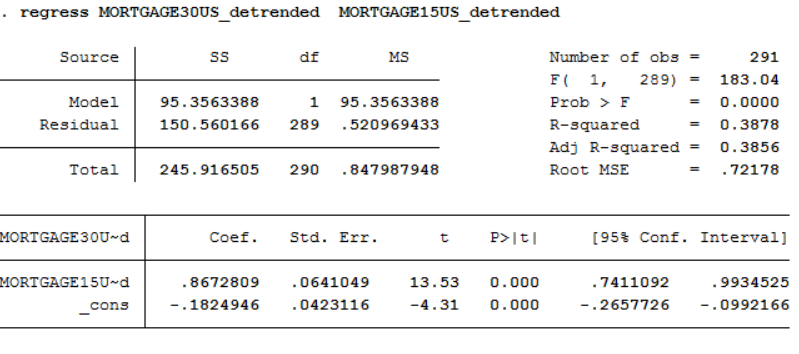
1. **Discussion of assumptions:**

Model C equation:

* SanFranciscoHomePriceIndex\_log\_detrended t = -755.21 -1.735 CPIAUCNS\_detrended +8.14 CSCICP03USM665S - 24.9 MORTGAGE30US\_detrended + 50.39 MORTGAGE15US\_detrended + 0.031 SP500\_detrended -5.01 Unemp -15.50 DLTIIT\_detrended

The classical assumptions for unbiased OLS are:

1. **TS1: Linear in parameters:** given the equation above, we can say that SanFranciscoHomePriceIndex\_log\_detrended t is a linear function of the 7 endogenous variables adjusted with their respective coefficients, as well as an intercept of -755.21.
2. **TS2: No perfect collinearity**: This assumption is worrisome. I suspect the variables MORTGAGE30US MORTGAGE15US to be collinear. I ran a regression on them:



This R squared of almost 40% show that they are related in some way. I will have to consider dropping on of them. Given that the 30 mortgage rate variable was non-significant in my models a) and b), I decided to drop that one.

1. **TS3: strict Exogeneity**: **E(**εt│X) = 0. The mean of the error term is unrelated to the values of the explanatory variable for all period.

This assumption could potentially fail. Indeed a value such as the index of the SP500 error term could be related to the value of that explanatory variable for the previous periods.

If a), b) and c) are respected, then the theorem for unbiaisness of OLS stands true.

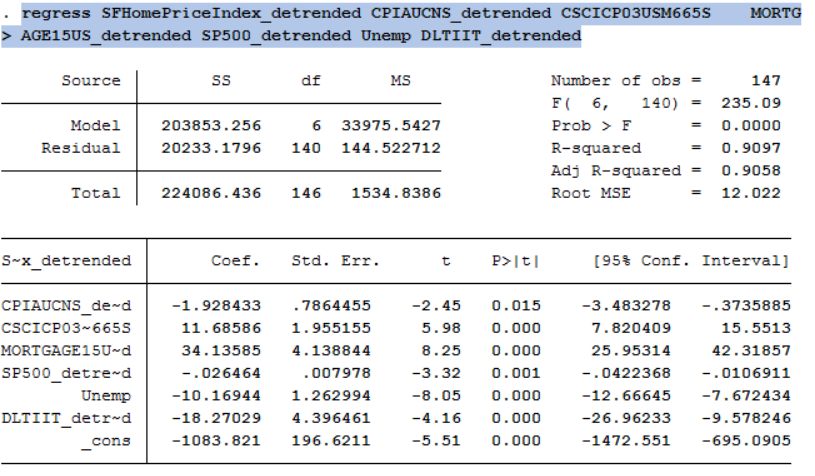
The classical assumptions for Consistent OLS are:

1. **TS1: Linear in parameters:** already taken care of in the unbiased section above.
2. **TS2: No perfect collinearity**: already taken care of in the unbiased section above.
3. **Contemporaneous exogeneity**:

This follow the same concept as strict exogeneity but is not as demanding. This could still be violated.

Note: Consistency matters more for very large datasets. Here, I do not believe we have enough observations for it to be relevant.

1. **Final model:**



Final equation of causal regression:

* SanFranciscoHomePriceIndex\_log\_detrended t = -1083.8 -1.92 CPIAUCNS\_detrended +11.64 CSCICP03USM665S + 34.13 MORTGAGE15US\_detrended - 0.026 SP500\_detrended -10.16 Unemp -18.27 DLTIIT\_detrended

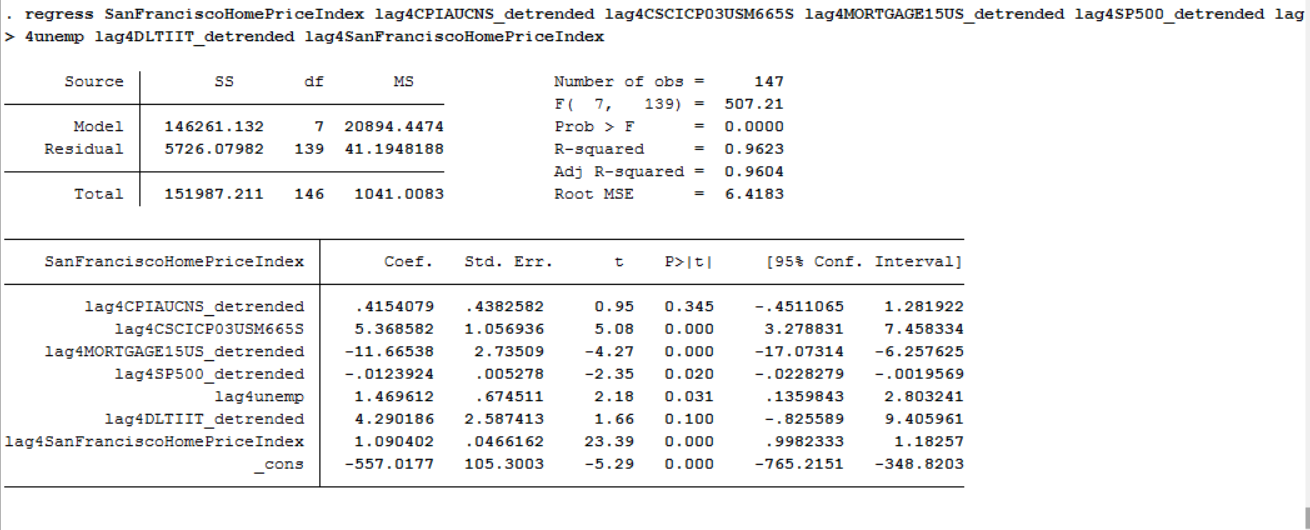
Adjusted R squared: 90%.

We removed the variable MORTGAGE30US\_detrended due to a collinearity issue.

Prob > F = 0.0000. The model is statically significant: null hypothesis is rejected.

1. **forecasting model**
2. **For February 2016:**

In order to forecast future unobserved values, I will use the regression I found in part IV) as it is the best indicator for the given SF home price index given the period t. For forecasting, we will also include lagged variables in the model.



This regression is calculated using lagged variables: the prefix lag4 corresponds to the variable lagged 4 time units (here months) this will be useful to calculate the February 2016 value given we have observations until November 2015. The adjusted R squared is very high, meaning the regression should be solid (96%)

The forecasting equation is:

* SanFranciscoHomePriceIndex t = -557 +0.415 lag4CPIAUCNS\_detrended + 5.37 lag4CSCICP03USM665S -11.66 lag4MORTGAGE15US\_detrended - 0.0124 lag4SP500\_detrended +1.47 lag4Unemp +4.29 lag4DLTIIT\_detrended + 1.09 SanFranciscoHomePriceIndex t-4

SanFranciscoHomePriceIndex t-4 is included to give a recursion to the curve and bring the r squared very high. The prediction will be moreaccurate.

Using this equation, we can plug in the real observed values for each variable:

SanFranciscoHomePriceIndex t where t equates to Feb 2016:

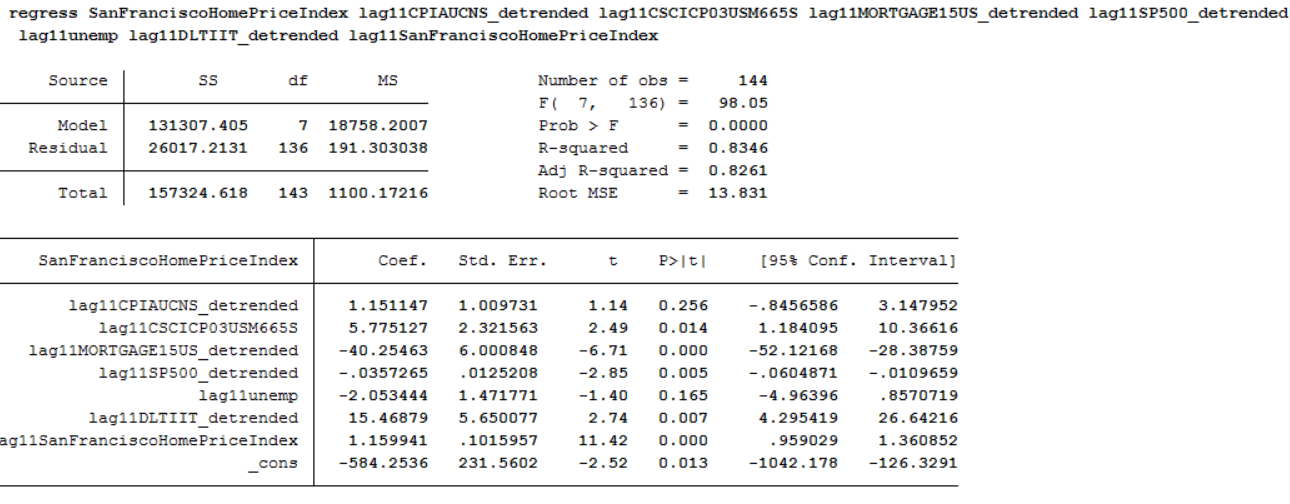
= -557+ -5.1220207 \*0.415 +5.37\*101.95 -11.66\* 0.039996 - 0.0124\*375.23099 +1.47\*5 +4.29\*0.53590816 + 1.09 \*218.4

=230.9317

Using this model, I predict the Sf home index to be around 230. Given that the index has a value of 218.4 as of November, this value seems realistic.

1. **For October 2016:**

Here, I will have to lag the variables 11 months in order to be able to predict the value of the SF home price index in October 2016.



This regression is also calculated using lagged variables: the prefix lag11 corresponds to the variable lagged 11 time units. The adjusted R squared is not quite as high as in the 4 months lagged example but still high (82%) I think we can still use this to get an idea of the prediction.

* SanFranciscoHomePriceIndex t = -584.25 +1.15 lag11CPIAUCNS\_detrended + 5.77 lag11CSCICP03USM665S -40.25 lag11MORTGAGE15US\_detrended - 0.0357 lag11SP500\_detrended -2.05 lag11Unemp +15.46 lag11DLTIIT\_detrended + 1.16 SanFranciscoHomePriceIndex t-11

Using this equation, we can plug in the real observed values for each variable:

SanFranciscoHomePriceIndex t where t equates to October 2016:

= -584.25 +1.15 \* -5.1220207 + 5.77 \*100.9513413687 -40.25 \* 0.039996- 0.0357 \* 375.23099 -2.05\* 5+15.46 \*0.5359082+ 1.16 \* 218.4

=228.7225

Again, the value of the estimate is in appearance consistent with a possible index value. We notice a slight decrease from February to October.

**Additional sources**

Moody’s case Shiller methodology: <http://www.moodysanalytics.com/~/media/Brochures/Economic-Consumer-Credit-Analytics/Examples/case-shiller-methodology.pdf>

Consumer price index handbook: <http://www.bls.gov/opub/hom/pdf/homch17.pdf>

S&P indices: <http://us.spindices.com/indices/>

Treasury yield information: <http://www.investopedia.com/terms/t/treasury-yield.asp>