

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.](#)

[Here is a link to the data compiled for this project.](#)

Method: Regression

OUTLINE:

I) **Introduction**

II) **Seasonality**

III) **explanatory variables**

IV) **Causal regressions:**

V) **forecasting model**

I) 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:

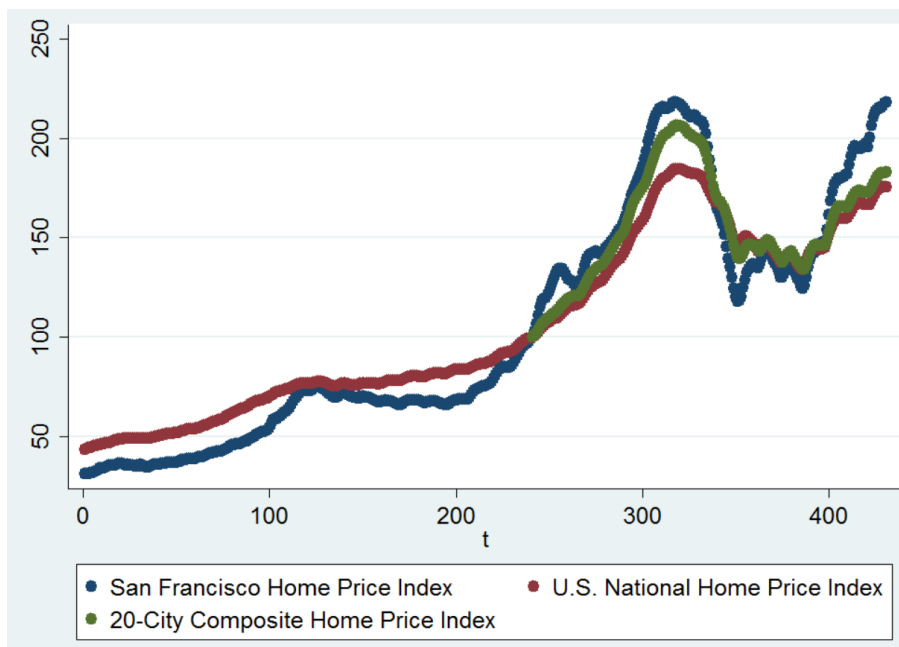
- Number of observation: 431.
- Each observation corresponds to a value for a given month.
- Time period range: from January 1980, until November 2015.
- Index Jan 2000=100

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 unbiasedness. The final step of our analysis will consist of a forecasting model, in order to generate potential values for February 2016 and October 2016.

II) 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:

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t, t = 1, 2, \text{etc.}$$

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. In 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.

```
. regress SanFranciscoHomePriceIndex t Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
```

Source	SS	df	MS	Number of obs =	431
Model	1139863.59	12	94988.6324	F(12, 418) =	134.14
Residual	295990.703	418	708.11173	Prob > F	= 0.0000
Total	1435854.29	430	3339.19603	R-squared	= 0.7939
				Adj R-squared	= 0.7879
				Root MSE	= 26.61

SanFrancis~x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
t	.413051	.0103052	40.08	0.000	.3927946 .4333075
Feb	-.3766621	6.272133	-0.06	0.952	-12.70551 11.95219
Mar	.1447313	6.272158	0.02	0.982	-12.18417 12.47363
Apr	1.229458	6.2722	0.20	0.845	-11.09953 13.55844
May	2.268074	6.27226	0.36	0.718	-10.06103 14.59717
Jun	2.883912	6.272336	0.46	0.646	-9.445339 15.21316
Jul	3.147805	6.272429	0.50	0.616	-9.181629 15.47724
Aug	2.947254	6.272539	0.47	0.639	-9.382396 15.2769
Sep	2.665314	6.272666	0.42	0.671	-9.664586 14.99521
Oct	2.211985	6.27281	0.35	0.725	-10.1182 14.54217
Nov	1.634212	6.272971	0.26	0.795	-10.69629 13.96471
Dec	.216356	6.316976	0.03	0.973	-12.20064 12.63335
_cons	13.53262	4.939409	2.74	0.006	3.823446 23.2418

- 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:

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

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.

2) Exponential trend

$$\log(y_t) = \alpha_0 + \alpha_1 t + \epsilon_t, t = 1, 2$$

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:

$$\text{SanFranciscoHomePriceIndex_log} = \log(\text{SanFranciscoHomePriceIndex})$$

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

```
. regress SanFranciscoHomePriceIndex_log t Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
```

Source	SS	df	MS	Number of obs =	431
Model	132.498837	12	11.0415697	F(12, 418) =	266.40
Residual	17.3248223	418	.041446943	Prob > F =	0.0000
				R-squared =	0.8844
				Adj R-squared =	0.8810
Total	149.823659	430	.348427114	Root MSE =	.20359

SanFrancis~g	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	.0044543	.0000788	56.50	0.000	.0042993	.0046092
Feb	-.003265	.0479856	-0.07	0.946	-.0975881	.0910581
Mar	-.0000613	.0479858	-0.00	0.999	-.0943848	.0942621
Apr	.0082015	.0479861	0.17	0.864	-.0861226	.1025256
May	.0173105	.0479865	0.36	0.718	-.0770145	.1116355
Jun	.0230017	.0479871	0.48	0.632	-.0713244	.1173278
Jul	.0258816	.0479878	0.54	0.590	-.068446	.1202091
Aug	.0246674	.0479887	0.51	0.608	-.0696618	.1189966
Sep	.022332	.0479896	0.47	0.642	-.0719991	.1166631
Oct	.0180217	.0479907	0.38	0.707	-.0763116	.112355
Nov	.0125977	.047992	0.26	0.793	-.081738	.1069334
Dec	.0079835	.0483286	0.17	0.869	-.087014	.102981
_cons	3.507485	.0377894	92.82	0.000	3.433204	3.581766

- 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:

```
regress SFHomePriceIndex_detrended t
```

Source	SS	df	MS	Number of obs = 431		
Model	5.8208e-11	1	5.8208e-11	F(1, 429) = 0.00		
Residual	296652.806	429	691.498382	Prob > F = 1.0000		
				R-squared = 0.0000		
				Adj R-squared = -0.0023		
Total	296652.806	430	689.890246	Root MSE = 26.296		

x_detrended	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	-4.67e-11	.0101806	-0.00	1.000	-.02001	.02001
_cons	2.62e-08	2.537717	0.00	1.000	-4.987905	4.987905

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

III) 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 years Mortgage rates

30-Year Fixed Rate Mortgage Average in the United States® (MORTGAGE30US)

Download Data

Source(s):	Freddie Mac
Release:	Primary Mortgage Market Survey
Units:	Percent Description of growth rate formulas
Frequency:	Monthly Aggregation Method: Average
Date Range:	1971-04-02 to 2016-01-21
File Format:	Excel
Seasonal Adjustment:	Not Seasonally Adjusted
Notes:	Data is provided "as is," by Freddie Mac® with no warranties of any kind, express or implied, including, but not limited to, warranties of accuracy or implied warranties of merchantability or fitness for a particular purpose. Use of the data is at the user's sole risk. In no event will Freddie Mac be liable for any damages arising out of or related to the data, including, but not limited to direct, indirect, incidental, special, consequential, or punitive damages, whether under a contract, tort, or any other theory of liability, even if Freddie Mac is aware of the possibility of such damages. Copyright, 2014, Freddie Mac. Reprinted with permission.
Updated:	2016-01-21 10:54 AM CST

Download Data

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:

. regress MORTGAGE30US t

Source	SS	df	MS	Number of obs = 431		
Model	4289.8952	1	4289.8952	F(1, 429) = 2681.31		
Residual	686.368727	429	1.5999271	Prob > F = 0.0000		
Total	4976.26393	430	11.5727068	R-squared = 0.8621		
				Adj R-squared = 0.8617		
				Root MSE = 1.2649		

MORTGAGE30US	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	-.0253571	.0004897	-51.78	0.000	-.0263196	-.0243946
_cons	13.70885	.1220668	112.31	0.000	13.46892	13.94877

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

2) 15 years Mortgage rates:

15-Year Fixed Rate Mortgage Average in the United States© (MORTGAGE15US)

Download Data

Source(s):	Freddie Mac
Release:	Primary Mortgage Market Survey
Units:	Percent <small>Description of growth rate formulas</small>
Frequency:	Monthly Aggregation Method: Average
Date Range:	1991-08-30 to 2016-01-21
File Format:	Excel
Seasonal Adjustment:	Not Seasonally Adjusted
Notes:	Data is provided "as is," with no warranties of any kind, express or implied, including, but not limited to, warranties of accuracy or implied warranties of merchantability or fitness for a particular purpose. Use of the data is at the user's sole risk. In no event will Freddie Mac be liable for any damages arising out of or related to the data, including, but not limited to direct, indirect, incidental, special, consequential, or punitive damages, whether under a contract, tort, or any other theory of liability, even if Freddie Mac is aware of the possibility of such damages. Copyright, 2014, Freddie Mac. Reprinted with permission.
Updated:	2016-01-21 10:56 AM CST

Download Data

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:

. regress MORTGAGE15US t

Source	SS	df	MS	Number of obs = 291		
Model	672.091006	1	672.091006	F(1, 289) = 1532.13		
Residual	126.773958	289	.438664215	Prob > F = 0.0000		
Total	798.864964	290	2.75470677	R-squared = 0.8413		
				Adj R-squared = 0.8408		
				Root MSE = .66232		

MORTGAGE15US	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	-.0180912	.0004622	-39.14	0.000	-.0190009	-.0171815
_cons	10.91732	.1377701	79.24	0.000	10.64616	11.18848

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

3) Unemployment rates:

Civilian Unemployment Rate (UNRATE)

Download Data	
Source(s):	US. Bureau of Labor Statistics
Release:	Employment Situation
Units:	Percent Description of growth rate formulas
Frequency:	Monthly
Date Range:	1948-01-01 to 2015-12-01
File Format:	Excel
Seasonal Adjustment:	Seasonally Adjusted
Notes:	<p>The unemployment rate represents the number of unemployed as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces.</p> <p>This rate is also defined as the U-3 measure of labor underutilization.</p> <p>The series comes from the "Current Population Survey (Household Survey)"</p> <p>The source code is: LNS14000000</p>
Updated:	2016-01-08 8:11 AM CST
Download Data	

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 reflection 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.

4) Consumer price index:

Consumer Price Index for All Urban Consumers: All Items (CPIAUCNS)

[Download Data](#)

Source(s):	US. Bureau of Labor Statistics
Release:	Consumer Price Index
Units:	Index 1982-1984=100 Description of growth rate formulas
Frequency:	Monthly
Date Range:	1913-01-01 to 2015-12-01
File Format:	Excel
Seasonal Adjustment:	Not Seasonally Adjusted
Notes:	Handbook of Methods - (http://www.bls.gov/opub/hom/pdf/homch17.pdf) Understanding the CPI: Frequently Asked Questions - (http://stats.bls.gov:80/cpi/cpifaq.htm)
Updated:	2016-01-20 9:21 AM CST

[Download Data](#)

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 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:

```
. regress CPIAUCNS t
```

Source	SS	df	MS	Number of obs = 431		
Model	919753.767	1	919753.767	F(1, 429) = .		
Residual	2310.21822	429	5.38512406	Prob > F = 0.0000		
Total	922063.985	430	2144.33485	R-squared = 0.9975		
				Adj R-squared = 0.9975		
				Root MSE = 2.3206		

CPIAUCNS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	.3712887	.0008984	413.27	0.000	.3695228	.3730545
_cons	82.43261	.2239471	368.09	0.000	81.99244	82.87278

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

5) 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:

```
. regress SP500 t
```

Source	SS	df	MS	Number of obs =	431
Model	112567941	1	112567941	F(1, 429) =	2718.30
Residual	17765385.1	429	41411.1541	Prob > F =	0.0000
				R-squared =	0.8637
				Adj R-squared =	0.8634
Total	130333326	430	303100.759	Root MSE =	203.5

SP500	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
t	4.107554	.0787834	52.14	0.000	3.952704 4.262403
_cons	-64.82661	19.6384	-3.30	0.001	-103.4261 -26.22715

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

6) Treasury Average Yield:

Treasury Inflation-Indexed Long-Term Average Yield (DLTIIT)

[Download Data](#)

Source(s):	Board of Governors of the Federal Reserve System (US)		
Release:	H.15 Selected Interest Rates		
Units:	Percent Description of growth rate formulas		
Frequency:	Monthly	Aggregation Method:	Average
Date Range:	2003-01-02	to	2016-01-20
File Format:	Excel		
Seasonal Adjustment:	Not Seasonally Adjusted		

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:

. regress DLTIIT t

Source	SS	df	MS	Number of obs =	155
Model	65.0078442	1	65.0078442	F(1, 153) =	277.40
Residual	35.8552435	153	.234347997	Prob > F	= 0.0000
				R-squared	= 0.6445
				Adj R-squared	= 0.6422
Total	100.863088	154	.654955115	Root MSE	= .4841

DLTIIT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
t	-.0144739	.000869	-16.66	0.000	-.0161907 -.012757
_cons	6.752335	.3100828	21.78	0.000	6.139738 7.364931

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

7) Consumer opinion confidence indicators

Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: OECD Indicator for the United States© (CSCICP03USM665S)

Download Data

Source(s):	Organization for Economic Co-operation and Development
Release:	Main Economic Indicators (Not a Press Release)
Units:	Normalised (Normal=100) Description of growth rate formulas
Frequency:	Monthly
Date Range:	1960-01-01 to 2015-03-01
File Format:	Excel
Seasonal Adjustment:	Seasonally Adjusted
Notes:	OECD descriptor ID: CSCICP03 OECD unit ID: IXNSA OECD country ID: USA All OECD data should be cited as follows: OECD, "Main Economic Indicators - complete database", Main Economic Indicators (database), http://dx.doi.org/10.1787/data-00052-en (Accessed on date) Copyright, 2014, OECD. Reprinted with permission.
Updated:	2015-06-08 2:13 PM CDT

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:

```
. regress CSCICP03USM665S t
```

Source	SS	df	MS	Number of obs = 423		
Model	9.39253789	1	9.39253789	F(1, 421) = 4.28		
Residual	923.927326	421	2.19460173	Prob > F = 0.0392		
Total	933.319864	422	2.21165845	R-squared = 0.0101		
				Adj R-squared = 0.0077		
				Root MSE = 1.4814		

CSCICP0~665S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	-.0012203	.0005899	-2.07	0.039	-.0023798	-.0000609
_cons	100.229	.1443139	694.52	0.000	99.94535	100.5127

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.

8) **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

IV) **Causal regressions:**

1) **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:

- Using Y= SanFranciscoHomePriceIndex, t included, X variables as they are.
- Using Y= SanFranciscoHomePriceIndex_log, t included, X variables as they are.

Method 2:

- Using Y= SFHomePriceIndex_detrended, X variables detrended
- Using Y= SFHomePriceIndex_log_detrended, X variables detrended

- Using Y= SanFranciscoHomePriceIndex, t included

```
. regress SanFranciscoHomePriceIndex t CPIAUCNS CSCICP03USM665S MORTGAGE15US MORTGAGE30US SP500 DLTIIT Unemp
```

Source	SS	df	MS	Number of obs = 147		
Model	128896.839	8	16112.1049	F(8, 138) = 126.90		
Residual	17522.0952	138	126.971705	Prob > F = 0.0000		
				R-squared = 0.8803		
				Adj R-squared = 0.8734		
				Root MSE = 11.268		
Total	146418.934	146	1002.86941			

SanFranciscoH~x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	1.231491	.3259778	3.78	0.000	.5869344	1.876049
CPIAUCNS	-2.33618	.7819566	-2.99	0.003	-3.882346	-.7900148
CSCICP03USM665S	6.97218	2.097486	3.32	0.001	2.824813	11.11955
MORTGAGE15US	35.4995	8.489016	4.18	0.000	18.71414	52.28486
MORTGAGE30US	-1.72116	11.7231	-0.15	0.883	-24.90129	21.45897
SP500	.0288105	.0162533	1.77	0.079	-.0033273	.0609483
DLTIIT	-20.87479	4.765712	-4.38	0.000	-30.29805	-11.45153
Unemp	-4.41948	1.77325	-2.49	0.014	-7.925732	-.9132268
_cons	-584.6029	264.6347	-2.21	0.029	-1107.866	-61.33985

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

b) Using Y= SanFranciscoHomePriceIndex_log, t included

```
. regress SanFranciscoHomePriceIndex_log t CPIAUCNS CSCICP03USM665S MORTGAGE15US MORTGAGE30US SP500 DLTIIT Unemp
```

Source	SS	df	MS	Number of obs = 147		
Model	4.6264368	8	.5783046	F(8, 138) = 135.16		
Residual	.590461955	138	.00427871	Prob > F = 0.0000		
				R-squared = 0.8868		
				Adj R-squared = 0.8803		
				Root MSE = .06541		
Total	5.21689875	146	.035732183			

SanFranciscoH~g	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	.0065244	.0018923	3.45	0.001	.0027827	.010266
CPIAUCNS	-.013186	.0045393	-2.90	0.004	-.0221615	-.0042105
CSCICP03USM665S	.0379934	.0121759	3.12	0.002	.0139179	.0620689
MORTGAGE15US	.1707353	.0492788	3.46	0.001	.0732961	.2681745
MORTGAGE30US	.0203534	.0680527	0.30	0.765	-.1142075	.1549143
SP500	.0002213	.0000944	2.35	0.020	.0000347	.0004079
DLTIIT	-.1198626	.027665	-4.33	0.000	-.1745647	-.0651605
Unemp	-.0251029	.0102937	-2.44	0.016	-.0454567	-.0047491
_cons	1.049918	1.536207	0.68	0.495	-1.98763	4.087465

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

c) Using Y= SFHomePriceIndex_detrended

```
. regress SFHomePriceIndex_detrended CPIAUCNS_detrended CSCICP03USM665S MORTGAGE30US_detrended MORTGAGE15US_detrended SP500_detrended Unemp DLTIIIT_detrended
```

Source	SS	df	MS	Number of obs = 147		
Model	205865.502	7	29409.3574	F(7, 139) = 224.35		
Residual	18220.934	139	131.085856	Prob > F = 0.0000		
				R-squared = 0.9187		
				Adj R-squared = 0.9146		
Total	224086.436	146	1534.8386	Root MSE = 11.449		

SFHomePriceIndex_det~d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
CPIAUCNS_detrended	-1.73482	.7506227	-2.31	0.022	-3.218934	-.2507051
CSCICP03USM665S	8.144688	2.069813	3.93	0.000	4.052301	12.23708
MORTGAGE30US_detrended	-24.95997	6.370618	-3.92	0.000	-37.55582	-12.36413
MORTGAGE15US_detrended	50.3977	5.724034	8.80	0.000	39.08026	61.71513
SP500_detrended	.0308778	.0164903	1.87	0.063	-.0017264	.063482
Unemp	-5.01018	1.783494	-2.81	0.006	-8.536464	-1.483896
DLTIIIT_detrended	-15.50108	4.246333	-3.65	0.000	-23.89684	-7.105329
_cons	-755.2191	205.1821	-3.68	0.000	-1160.901	-349.5377

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

d) Using Y= SFHomePriceIndex_log_detrended

```
. regress SFHomePriceIndex_log_detrended CPIAUCNS_detrended CSCICP03USM665S MORTGAGE30US_detrended MORTGAGE15US_detrended SP500_detrended Unemp DLTIIIT_detrended
```

Source	SS	df	MS	Number of obs = 147		
Model	11.74713	7	1.67816142	F(7, 139) = 297.27		
Residual	.784692955	139	.005645273	Prob > F = 0.0000		
				R-squared = 0.9374		
				Adj R-squared = 0.9342		
Total	12.5318229	146	.085834404	Root MSE = .07514		

SFHomePriceIndex_log~d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
CPIAUCNS_detrended	-.0031605	.0049259	-0.64	0.522	-.0128999	.0065789
CSCICP03USM665S	.0575407	.013583	4.24	0.000	.0306847	.0843967
MORTGAGE30US_detrended	-.3670691	.0418067	-8.78	0.000	-.4497284	-.2844098
MORTGAGE15US_detrended	.4191084	.0375636	11.16	0.000	.3448386	.4933782
SP500_detrended	.0002558	.0001082	2.36	0.019	.0000418	.0004697
Unemp	-.0349507	.011704	-2.99	0.003	-.0580916	-.0118097
DLTIIIT_detrended	-.0302758	.0278663	-1.09	0.279	-.0853723	.0248208
_cons	-5.307202	1.346493	-3.94	0.000	-7.969457	-2.644947

- 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
- DLTIT_detrended, P-value of 0.279

2) 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 DLTIT_detrended

The classical assumptions for unbiased OLS are:

- 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.
- TS2: No perfect collinearity:** This assumption is worrisome. I suspect the variables MORTGAGE30US MORTGAGE15US to be collinear. I ran a regression on them:

```
. regress MORTGAGE30US_detrended MORTGAGE15US_detrended
```

Source	SS	df	MS	Number of obs = 291		
Model	95.3563388	1	95.3563388	F(1, 289) = 183.04		
Residual	150.560166	289	.520969433	Prob > F = 0.0000		
Total	245.916505	290	.847987948	R-squared = 0.3878		
				Adj R-squared = 0.3856		
				Root MSE = .72178		

MORTGAGE30U~d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
MORTGAGE15U~d	.8672809	.0641049	13.53	0.000	.7411092	.9934525
_cons	-.1824946	.0423116	-4.31	0.000	-.2657726	-.0992166

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.

c) **TS3: strict Exogeneity:** $E(\epsilon_t | X) = 0$. 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 unbiasedness of OLS stands true.

The classical assumptions for Consistent OLS are:

- a) **TS1: Linear in parameters:** already taken care of in the unbiased section above.
- b) **TS2: No perfect collinearity:** already taken care of in the unbiased section above.
- c) **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.

3) Final model:

```
. regress SHomePriceIndex_detrended CPIAUCNS_detrended CSCICP03USM665S MORTG
> AGE15US_detrended SP500_detrended Unemp DLTIIIT_detrended
```

Source	SS	df	MS	Number of obs =	147
Model	203853.256	6	33975.5427	F(6, 140) =	235.09
Residual	20233.1796	140	144.522712	Prob > F =	0.0000
				R-squared =	0.9097
				Adj R-squared =	0.9058
Total	224086.436	146	1534.8386	Root MSE =	12.022

S~x_detrended	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
CPIAUCNS_de~d	-1.928433	.7864455	-2.45	0.015	-3.483278	-.3735885
CSCICP03~665S	11.68586	1.955155	5.98	0.000	7.820409	15.5513
MORTGAGE15U~d	34.13585	4.138844	8.25	0.000	25.95314	42.31857
SP500_detre~d	-.026464	.007978	-3.32	0.001	-.0422368	-.0106911
Unemp	-10.16944	1.262994	-8.05	0.000	-12.66645	-7.672434
DLTIIIT_detr~d	-18.27029	4.396461	-4.16	0.000	-26.96233	-9.578246
_cons	-1083.821	196.6211	-5.51	0.000	-1472.551	-695.0905

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 DLTIT_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.

V) forecasting model

1) 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.

```
. regress SanFranciscoHomePriceIndex lag4CPIAUCNS_detrended lag4CSCICP03USM665S lag4MORTGAGE15US_detrended lag4SP500_detrended lag4unemp lag4DLTIIT_detrended lag4SanFranciscoHomePriceIndex
```

Source	SS	df	MS	
Model	146261.132	7	20894.4474	Number of obs = 147
Residual	5726.07982	139	41.1948188	F(7, 139) = 507.21
Total	151987.211	146	1041.0083	Prob > F = 0.0000
				R-squared = 0.9623
				Adj R-squared = 0.9604
				Root MSE = 6.4183

SanFranciscoHomePriceIndex	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lag4CPIAUCNS_detrended	.4154079	.4382582	0.95	0.345	-.4511065 1.281922
lag4CSCICP03USM665S	5.368582	1.056936	5.08	0.000	3.278831 7.458334
lag4MORTGAGE15US_detrended	-11.66538	2.73509	-4.27	0.000	-17.07314 -6.257625
lag4SP500_detrended	-.0123924	.005278	-2.35	0.020	-.0228279 -.0019569
lag4unemp	1.469612	.674511	2.18	0.031	.1359843 2.803241
lag4DLTIIT_detrended	4.290186	2.587413	1.66	0.100	-.825589 9.405961
lag4SanFranciscoHomePriceIndex	1.090402	.0466162	23.39	0.000	.9982333 1.18257
_cons	-557.0177	105.3003	-5.29	0.000	-765.2151 -348.8203

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 more accurate.

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.

2) 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.

```
regress SanFranciscoHomePriceIndex lag11CPIAUCNS_detrended lag11CSCICP03USM665S lag11MORTGAGE15US_detrended lag11SP500_detrended lag11unemp lag11DLTIIT_detrended lag11SanFranciscoHomePriceIndex
```

Source	SS	df	MS	Number of obs = 144			
Model	131307.405	7	18758.2007	F(7, 136)	=	98.05	
Residual	26017.2131	136	191.303038	Prob > F	=	0.0000	
				R-squared	=	0.8346	
				Adj R-squared	=	0.8261	
Total	157324.618	143	1100.17216	Root MSE	=	13.831	

SanFranciscoHomePriceIndex	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lag11CPIAUCNS_detrended	1.151147	1.009731	1.14	0.256	-.8456586	3.147952
lag11CSCICP03USM665S	5.775127	2.321563	2.49	0.014	1.184095	10.36616
lag11MORTGAGE15US_detrended	-40.25463	6.000848	-6.71	0.000	-52.12168	-28.38759
lag11SP500_detrended	-.0357265	.0125208	-2.85	0.005	-.0604871	-.0109659
lag11unemp	-2.053444	1.471771	-1.40	0.165	-4.96396	.8570719
lag11DLTIIT_detrended	15.46879	5.650077	2.74	0.007	4.295419	26.64216
lag11SanFranciscoHomePriceIndex	1.159941	.1015957	11.42	0.000	.959029	1.360852
_cons	-584.2536	231.5602	-2.52	0.013	-1042.178	-126.3291

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

$$\text{lag11SP500_detrended} - 2.05 \text{ lag11Unemp} + 15.46 \text{ lag11DLTIT_detrended} + 1.16 \text{ SanFranciscoHomePriceIndex}_{t-11}$$

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

$\text{SanFranciscoHomePriceIndex}_t$ 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.moodyanalytics.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>