July 05, 2017

Economics Group

WELLS FARGO SECURITIES

Special Commentary

John E. Silvia, Chief Economist
john.silvia@wellsfargo.com • (704) 410-3275
Azhar Iqbal, Econometrician
azhar.iqbal@wellsfargo.com • (704) 410-3270
Michael Pugliese, Economic Analyst
michael.d.pugliese@wellsfargo.com • (704) 410-3156

Big Data Applications in the Economics/Financial World Part II: Econometric Modeling in the 21st Century

"One of the first things taught in introductory statistics books is that correlation is not causation.
It is also one of the first things forgotten." – Thomas Sowell

Executive Summary

Data analysis and econometric modeling are major tools that help decision makers design effective polices. This ever evolving world is driven by continuous developments in econometric tools and data sources. One such development in recent years is the use of big data—large and complex datasets that present both opportunities and challenges for analysts and their statistical toolkits. To inform our readers about the potential benefits and limitations of big data applications, we started a two-part series, and the first report looked at the potential benefits from big data for analysts and decision makers.[1]

This report focuses on issues and solutions related to econometric modeling and forecasting using big data. It is worth mentioning that we recognize that there are other issues, such as privacy and cyber security concerns, that represent huge hurdles for the advancement of utilizing big data on a mass scale. We are not lawyers, information technology or privacy experts, however, and as such, we will solely focus on using big data for modeling and analysis from an economics perspective.

This report focuses on issues and solutions related to econometric modeling and forecasting using big data.

Data as Insight

Typically, analysts utilize a data series to represent or proxy a sector's (or the overall economy's) activities, and by analyzing that series an analyst gains insight about the sector's state. For example, the Bureau of Economic Analysis releases personal spending data every month, and analysts use that series to learn about consumers' behavior and its impact on the overall economy, Figures 1 & 2. This data lags by about one month, however, and is only available in the aggregate. More granular data that include other useful information, such as demographic data, are often unavailable in real-time. Big data, with its hundreds of millions of observations, high frequency and detailed richness, can help to fill this information void. These very traits, however, can create both known and unknown challenges for analysts using traditional statistical techniques and applications. The relative newness of big data analysis likely means that as this phenomenon matures over time, problems with current methods will arise and new statistical tools and techniques will be needed to address these currently unknown problems lurking in the vast sea of big data. For now, however, we cannot know these unknown unknowns, and we must rely on the tools at hand. This report outlines some known potential problems and suggests a few key statistical principles, methods and best practices to bear in mind when conducting big data analysis.

The dream of finding a crystal ball is not new, and some think big data may be the tool that makes this dream a reality, but we do not think big data constitutes a perfect crystal ball. A few months

[1] "Big Data Applications in the Economics/Financial World Part I: Opportunities and Challenges," published on April 6, 2017. The report is available upon request.

ago, Vice-Chairman Fischer shared one example from his undergraduate years in the 1960s.[2] Fischer said that one of his friends told him that soon researchers would be able to build a model that could predict the future course of the economy accurately. Over 50 years have passed, and we are still waiting for Godot.[3]

Figure 1

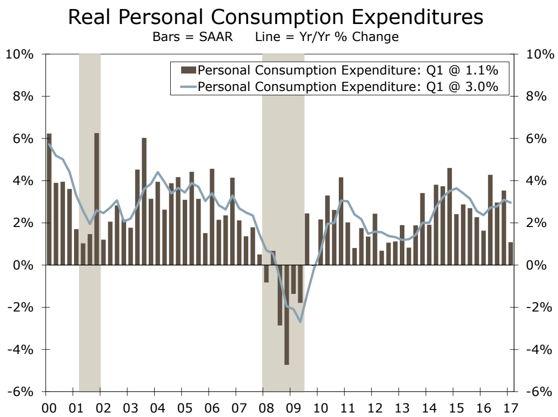
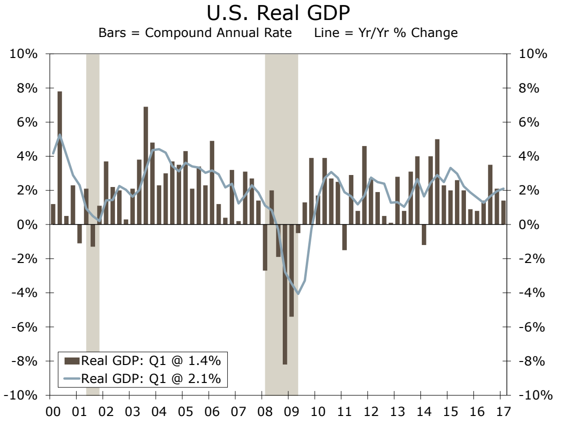


Figure 2



Source: U.S. Department of Commerce and Wells Fargo Securities

An effective analyst chooses tools according to the problem at hand, not simply because everyone else is using that tool.

The economy is a complex and evolving system, and these characteristics make it hard for analysts to predict its future path accurately. Almost every econometric/statistical method contains some set of assumptions, and if those assumptions are violated then the results will not be reliable. In addition, data is a crucial element of a model, and if the data is not reliable then it is almost impossible to obtain a reliable result (a "garbage-in garbage-out" problem).[4] Raw big data tend be very volatile and potentially face non-stationary and seasonality issues, which can cause misleading results. There are some other potential issues posed by big data for modeling, and we discuss those problems and potential solutions in this report.

In short, big data has the potential to improve modeling and decision making significantly. However, big data has issues, and those issues could mislead analysts if not addressed. Big data is another tool in the arsenal for decision makers, but an effective analyst chooses tools according to the problem at hand, not simply because everyone else is using that tool. In our view, at present, an abundance of information, not lack of information, is an underappreciated problem for modeling. We need to excerpt relevant and useful information and then utilize that information in the modeling process to design effective policies. An important task for analysts is to remove irrelevant noise from the data/information, a potential challenge in a world of exponentially growing data availability. We provide a framework that helps decision makers extract useful information and then effectively utilize that information in the decision making process.

Lots of Small Problems with Big Data: Is Being Rich Enough?

Big data applications are growing at an unprecedented speed.[5] We believe, as mentioned in our first report, that big data will enhance our understanding of economic agents' behavior and that better modeling will lead to more effective decision making. However, there are lots of small problems with big data and those small problems, if not resolved, can create a big mess. Before we discuss those problems and potential solutions, we want to share an insightful example. The example is not recent but rather from 1936. The concept of big data may not be as "new" as people

[2] Complete speech by Fischer is available at the following link:
https://www.federalreserve.gov/newsevents/speech/files/fischer20170211a.pdf
[3] "Waiting for Godot" is a play by Samuel Beckett about waiting for someone or something that never arrives https://en.wikipedia.org/wiki/Waiting_for_Godot
[4] Silvia, John E., Azhar Iqbal, Sam Bullard, Sarah Watt and Kaylyn Swankoski. (2014). *Economic and Business Forecasting: Analyzing and Interpreting Econometric Results*. Wiley, 2014.
[5] For more detail see Das (2016), available at:
http://www.imf.org/external/pubs/ft/fandd/2016/09/das.htm

believe as there has always been a desire to include full information in models that would provide a more accurate forecast.

Harford (2014) provided an interesting example about the 1936 U.S. election.[6] For researchers, predicting an election's outcome has always been an area of strong interest. In 1936, the Literary Digest decided to use "big data" to predict whether the Republican Alfred Landon or the Democrat President Franklin Delano Roosevelt would win the election. The Literary Digest's sample consisted of 10 million people, as they planned to reach out to 10 million people or roughly a quarter of voters at the time. Around 2.4 million people responded to the Literary Digest survey and, after calculating and rechecking estimates, the Digest predicted a comfortable victory for Landon by 55 percent to 41 percent. The actual election result was completely different than the Digest's prediction, and Roosevelt enjoyed a comfortable victory of 61 percent to 37 percent. Missing the election's outcome was not the only embarrassment faced by the Literary Digest: the other, which probably hurt even more, was that George Gallup, the opinion poll pioneer, predicted the election's outcome correctly by conducting interviews with only about 3,000 people.

In 1936, the Literary Digest decided to use "big data" to predict who would win the presidential election; the result was completely different than the Digest's prediction.

If You Want an Accurate Prediction, Take Sampling Bias Seriously

Does that mean big data (10 million people) is useless? Or that small data (3,000 interviews) is good? How is that example relevant for today's big data applications? Before we answer these questions, we take a step back and ask a very basic question: why do we collect data? Analysts need data to understand a situation and make informed decisions about the future, such as the Literary Digest wanting to predict the 1936 election's outcome. Analysts utilize data and build models to make predictions. In the present example, it may not be possible to ask every voter about their preferences and for whom they are going to vote. It is likely very costly to reach out to even a simple majority of voters. So, analysts choose a sample, collect data and then utilize that data to make predictions. In Statistics 101, we have a population (in the present case, all registered voters in 1936 is considered the population) and a sample (the sample for the Literary Digest is 10 million and 3,000 for Gallup).[7] In statistics/econometrics, analysts almost always utilize sample data and then generalize the results to the population as a whole. Statistically, a good sample is expected to mirror the population from which it comes.

So what went wrong with the Literary Digest's prediction? Or did Gallup just get lucky? When it comes to data analysis, the sample size is not the most important element and a small sample can sometimes provide more accurate results. The most important thing is that the sample must be random (unbiased). In other words, a sample must be a good representative of the population: a small, random sample can provide more accurate results than a large biased sample in statistical analysis. In the polling example, Gallup's small sample was random and the Literary Digest's "big data" was a biased sample. As Harford (2014) explains, the Literary Digest selected 10 million people using automobile registrations and telephone directories and, in 1936, mostly prosperous people had one or both of those luxury items. Furthermore, a majority of those selected well-off people were Republicans. Gallup, on the other hand, randomly selected from a well representative sample of 3,000 people. Therefore, a biased sample including millions of responses generated an incorrect prediction. There are several other examples in the Harford (2014) study where biased samples provided misleading results.

Simply increasing sample size is no guarantee of accurate results.

What we are suggesting using this insightful example is that when we are selecting a sample, we should consider several things beyond just size. For example, in the above example, factors such as gender, region, location (rural versus urban), income group, ethnicity and other factors should be considered when drawing a sample in order to obtain reliable results. Randomly selecting from the established sample is then the crucial next step. Simply increasing sample size is no guarantee of accurate results. Unfortunately, some analysts fall into the trap of thinking that, by having so much data, their results will automatically be more accurate. The problem gets worse when a

[6] There are many other interesting examples in Harford, T. (2014). Big Data: Are we making a big mistake? *Financial Times*: https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0
[7] Since the Literary Digest selected and reached out to 10 million people, the sample is 10 million and not the 2.4 million who responded the survey.

researcher assumes N=all, or rather that he/she has information for the whole population (in other words complete information). There are several examples of such studies in Harford (2014).

Figure 3

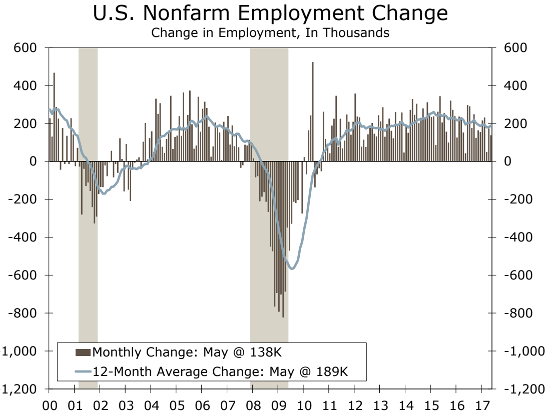
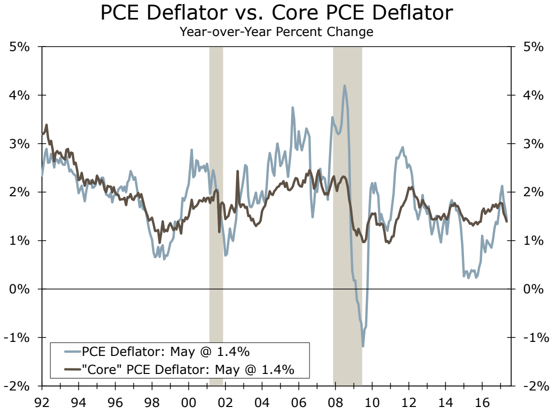


Figure 4



Source: U.S. Department of Labor, U.S. Department of Commerce and Wells Fargo Securities

Big Data Adds "Fun" to Time Series Analysis[8]

We re-ask the question about sampling bias and its relevance with current big data applications. Furthermore, we add one more question: what happens to the sampling bias if we are using time series (or panel data) instead of cross section? Most macroeconomic data, such as nonfarm payrolls (Figure 3) and the PCE deflator (Figure 4) are survey-based, time series measures, and a well representative sample is essential for accurate results. In other words, the dangers of sampling bias are present for all types (cross section, time series and panel data for example) of data series. Current big data applications provide some great insights for analysts if done correctly. However, these applications pose serious issues if we ignore certain statistical rules.

A recent trend used social networking data for analysis, but this only provides data on the population that uses these platforms.

A recent trend is to use social networking data (Twitter/Google trends etc.) to forecast macroeconomic/equity prices, Nyman et. al (2014).[9] Some of these studies claim since they have access to all users' data (all Twitter users' data for example) that thereby their N=all (information about whole population). In reality, they only have data from those people who have Twitter accounts, and not all economic/financial agents use Twitter. Furthermore, some users are more active than others. Again, analysts must be wary of being too overzealous when faced with the exciting possibilities brought on by big data.

A Statistical Framework to Bake the Big Data Cake

The sampling error/bias is not the only problem an analyst faces when dealing with big data. Even if the data are obtained using a good sample, there still might be issues related to that dataset. Moreover, if we ignore these issues, the model might produce unreliable results. Before we discuss the issues and challenges posed and potential solutions to those problems, we take a brief look at the modern history of econometric modeling. We believe a quick "history-tour" would help

[8] Typically, in econometrics, there are three types of data: cross section, time series and panel data. A data series collected for many objects/individuals at the same time is called cross section. A good example is the Literary Digest survey of 10 million people. A time series data represents an entity's behavior over time and Figure 1 is a time series of the U.S. private consumption. The panel (or sometimes known as longitudinal) data refers to a dataset of the same objects or individuals for multiple times. A good example of a panel dataset is personal income of all 50 states for the last 30 years. For more detail about different data types see, Greene, W. (2011). *Econometric Analysis*. 7th edition, Pearson

**[9] See Nyman et al (2014) for more examples;
https://www.ecb.europa.eu/events/pdf/conferences/140407/TuckettOrmerod_BigDataAndEconomicForecastingATop-DownApproachUsingDirectedAlgorithmicTextAnalysis.pdf**

our readers to understand the problems posed by today's big data and solutions to those problems.[10]

The first Nobel Prize in Economics was awarded in 1969 to Ragnar Frisch and Jan Tinbergen "for having developed and applied dynamic models for the analysis of economic processes."[11] In 1936, Jan Tinbergen built the first empirical macro econometric (or macroeconomic) model for the Netherlands and later he built similar models for the United States and the United Kingdom.[12] One common issue with empirical time series models is autocorrelation, and in the presence of autocorrelation model results may not be reliable.[13] Furthermore, Durbin and Watson (1950) developed the first formal way of testing for autocorrelation.[14]

Essentially, the first empirical model was built in 1936 and at that time there was no statistical test to identify autocorrelation. Furthermore, Granger and Newbold (1974) provided a statistical proof that if underlying data are non-stationary, then OLS results are spurious and we tend to find a strong statistical relationship between variables of interest, even if there is none.[15] Almost every time series technique assumes the underlying dataset is stationary and if the data are non-stationary then results are misleading, see Silvia and Iqbal et. al (2014) for a detailed discussion about stationary, non-stationary and spurious results.[16] Dickey and Fuller (1979) introduced the first statistical test to verify whether a variable is stationary or non-stationary.[17] There are some other important problems that can cause spurious results such as the ARCH effect and structural breaks. To save space, we are not going to discuss those issues in detail and we suggest Silvia and Iqbal et. al (2014) for interested readers.

Basically, after the first empirical model of Tinbergen (1936), researchers spent more than 40 years to develop a test (Dickey and Fuller, 1979) to verify whether the data are stationary and hence results are reliable. How do these statistical issues and the brief history of time series modeling relate to big data modeling in the present day? In our view, the brief history of time series modeling along with the related statistical issues are directly related with the today's big data modeling. Furthermore, analysts who use big data can learn from the past and avoid mistakes such as assuming a dataset is stationary without verifying and dealing with the spurious regression problem. Before Tinbergen's 1936 macroeconomic model, most people were involved either in theoretical research (without using real world data) or using cross section data. One major reason for a lack of time series analysis at the time was the lack of availability of time series data. Most countries regularly started collecting macroeconomic data (such as GDP, CPI etc.) after World War II. Therefore, the time series modeling of the 1930s was a big deal, akin to the big data applications of the past decade or so.

The time series modeling of the 1930s was a big deal, akin to the big data applications of the past decade or so.

In fairness to past researchers, if an analyst is dealing with cross section data then she should not worry about autocorrelation and non-stationary issues, as these problems are unique to time

[10] It is important to note that whenever an analyst builds a model using any kind of data, whether big data or traditional, there are several issues that need to be addressed. Otherwise, a model's results won't be reliable. This report focus on time series data to highlight several statistical issues and solution to those problems. Readers can see Greene's (2011) book for issues related to cross section and panel data.

[11] https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1969/

[12] For a history of macro econometric modeling see Dhaene G. and Barten P. (1989). When it all began. *Economic Modelling*, Vol 6, no 2.

[13] Autocorrelation presence is an indication that important information (important variables and/or correct function form of the variables) is missing. Furthermore, if irrelevant information (irrelevant variables and/or wrong functional form of the relevant variables) is included in the model that would also cause autocorrelation. For more detail about autocorrelation and its consequences to results see Greene (2011).

[14] Durbin J and Watson S. (1950). Testing for serial correlation in least squares regression, I. *Biometika*, vol 37 no 3-4.

[15] Granger C and Newbold P. (1974). Spurious Regression in Econometrics. *Journal of Econometrics*, Vol 2, pp111-120.

[16] Silvia, John E., Azhar Iqbal, Sam Bullard, Sarah Watt and Kaylyn Swankoski. (2014). *Economic and Business Forecasting: Analyzing and Interpreting Econometric Results*. Wiley, 2014.

[17] If the mean and variance of a variable is constant over time, then that variable is stationary, see Silvia and Iqbal et.al (2014) or Dickey, D. and Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of American Statistical Association*, vol 74, pp427-431.

series data. However, cross section data have their own problems, such as heteroskedasticity which raises questions about the reliability of results, see Greene (2011) for more detail about cross section data issues. What we are suggesting is that whenever we utilize new data/modeling techniques, we should not just concentrate on the benefits of the new tools but also consider the potential issues/problems of those new tools in order to obtain accurate results/analysis.

Big data modeling/analysis consists of time series data, and as such we must test for non-stationarity, the ARCH-effect, structural breaks and the potential for autocorrelation. Big data applications are a relatively new territory and therefore statistical issues related with big data fall into three categories (1) known-knowns, (2) known-unknowns and (3) unknown-unknowns. Known-known issues include making sure to test for standard time series issues, such as autocorrelation. Known-unknown issues include challenges such as seasonality, because most macroeconomic data are seasonally adjusted and big data might need to be adjusted for regular noise. Spending patterns throughout the seasons can vary, holidays can increase the volatility of the data, and certain events can create confusing noise, such as the release of a hotly anticipated new smart phone. Big data analysts need to handle seasonal patterns otherwise seasonality can influence the results and provide misleading conclusions.

Figure 5

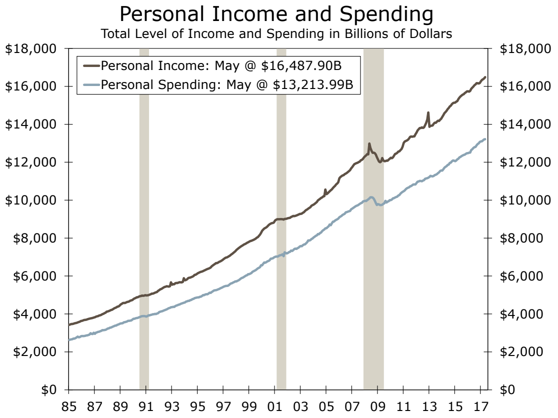
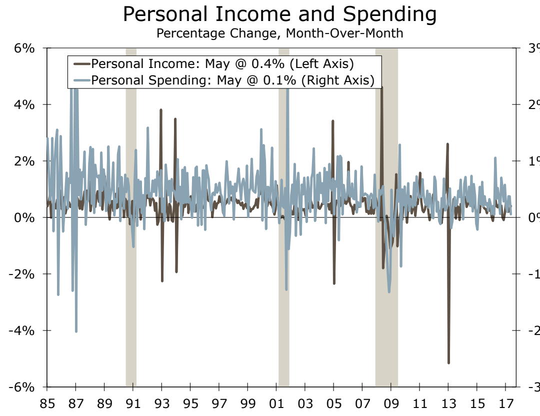


Figure 6



Source: U.S. Department of Commerce and Wells Fargo Securities

There are lots of unknown-unknown issues in big data applications, in our view.

There are lots of unknown-unknown issues in big data applications, in our view. For example, traditional macroeconomic data are released at a fixed frequency (monthly unemployment rate and quarterly GDP, for example) and are based on a fixed number of respondents. Big data, such as credit/debit cards transactions, may not have a fixed frequency or the same number of respondents. It is likely that some days will have more transactions than others and more people will be using credit/debit cards now compared to 10 years ago. Furthermore, big data consist of hundreds of millions (sometimes billions) of observations from (at least) millions of respondents, and these features would create significantly higher volatility that the traditional dataset and that volatility would influence results.

The biggest unknown-unknown, in our view, is whether the traditional econometric/statistical tools (such as tests for autocorrelation/non-stationarity and structural break) can handle so many observations from different respondents and provide accurate results. Future research should provide an answer to this question, just as the first time series macroeconomics model was developed in 1936 and the first non-stationarity test was introduced more than 40 years later in 1979. We believe in the near future researchers will find appropriate tools to handle big data, but the challenges today remain.

Three Words to Obtain Accurate Results: Test, Test, Test

In research, techniques are often "innocent until proven guilty," and as such we rely on traditional statistical tests for stationarity and structural breaks until researchers find issues with those tests. The first step of big data analysis using time series data should be to verify the data are stationary so we can avoid spurious regression problems. We use personal income and spending data in levels (Figure 5) and month-over-month percent change (Figure 6) to show the spurious regression problem. The results are reported in Tables 1 & 2, see Appendix for all results. Using the level form of both personal income and spending variables, we found a very strong relationship (R-squared is 0.9986), but the non-stationarity test (ADF test) shows both series are non-stationary at level form, Table 1. The month-over-month form of both series is stationary and the relationship is very weak (R-squared drops significantly to only 0.0309, Table 2) compared to the level form estimates. The level form of personal income and spending contains a deterministic trend (both have a clear upward trend over time) and that is why both series are non-stationary. Moreover, usually, variables with a clear trend produce a very high R-squared and t-value (signs of a strong statistical relationship) even if there is no relationship between the variables. Therefore, analysts dealing with time series big data must check for stationarity in the variables of interest to obtain reliable results.

Figure 7

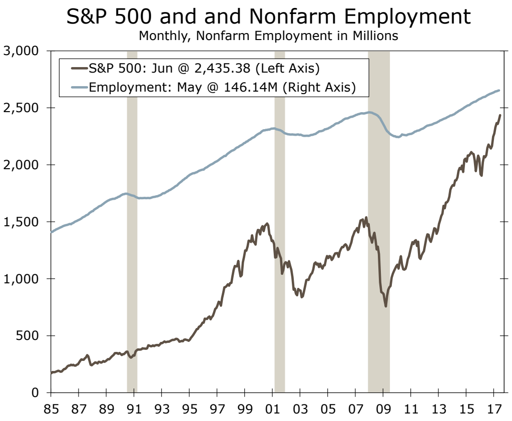
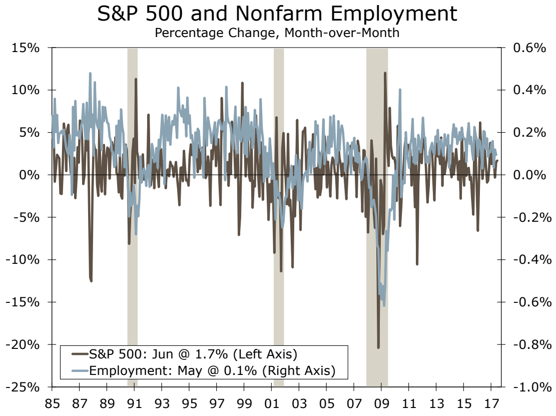


Figure 8



Source: U.S. Department of Labor, Standard & Poor's and Wells Fargo Securities

The next important issue to consider is the volatility of big data because, if not handled properly, volatility can lead to misleading results. As mentioned earlier, big data consists of hundreds of millions or billions of observations from millions of respondents/users, and naturally a dataset based on large number of observations/respondents tends to be very volatile. Therefore, we must test whether the volatility of the data can influence results. The Autoregressive Conditional Heteroscedasticity (ARCH) approach helps us to test the volatility issue and provide reliable results in the case of a volatile dataset.[18] We use the S&P 500 Index and nonfarm payrolls data in levels (Figure 7) and month-over-month percent change, MoM, (Figure 8) to show the volatility problem.

The ARCH approach can help us to test the volatility issue inherent in big data.

Using the level of both variables, we found a very strong relationship (in terms of R-squared), but because the level form of both variables is non-stationary the results are spurious. A major reason we use the level form of both variables is to show another common mistake in regression analysis by some analysts. That is, both series have different measurement scales (the S&P 500 is an index and employment data are reported in thousands) and the interpretation of the estimated coefficient is not meaningful, as a one unit (one thousand) increase in employment is associated with 0.039 unit increase in the S&P 500 index. The MoM form is stationary, and we can explain the coefficient in a more meaningful way, such as a one percentage point increase in employment is associated with a 3.119 percentage point increase in the S&P 500 index. Although the MoM

[18] Engle (1982) introduced the Autoregressive Conditional Heteroscedasticity (ARCH) model to test for volatility clusters and financial data often exhibits a volatility cluster. When some periods are more volatile than others, this is called a volatility cluster in simple words. If variables exhibit volatility clusters, then OLS results are not reliable and we need to employ an ARCH model. For more detail about ARCH and volatility cluster, see Silvia and Iqbal et. al (2014) and Engle R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation. *Econometrica*, Vol 50 (4).

form of both employment and the S&P 500 index is stationary (Table 1), we need to verify the volatility cluster (or ARCH effect). The results suggest both series have a volatility problem (the ARCH effect), and we need to utilize ARCH modeling to obtain reliable results.[19] The results based on the ARCH model are significantly different than those provided by OLS (using MoM form) and lead to different conclusions. The OLS MoM form suggests a statistically significant relationship between employment and S&P 500 growth rates; however, the ARCH/GARCH method suggests that there is no relationship between these two variables. Therefore, analysts must test for the volatility cluster and, if the model has a volatility problem, then analysts can utilize the ARCH approach to obtain reliable results. We believe, given the size of big data, most models would have volatility problems and would need to correct for this volatility issue. We believe ARCH is the best tool for the job to obtain reliable results.

Figure 9

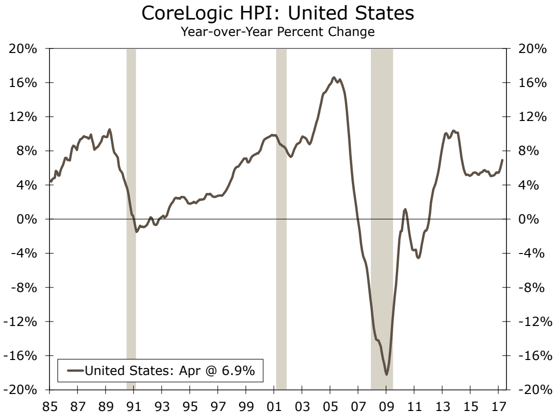
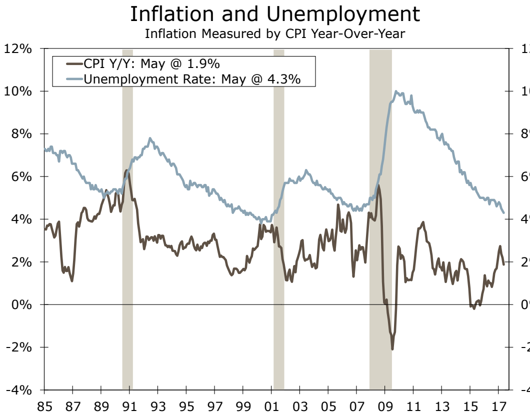


Figure 10



Source: CoreLogic, U.S. Department of Labor and Wells Fargo Securities

We live in an evolving world where everything is changing, such as consumers' preferences, policies, technologies and so on. These changes affect the underlying relationship between different variables and thus we need to test for structural breaks in our variables of interest. A structural break indicates that the series' behavior is different for the post-break era than those of the pre-break period. Furthermore, due to the different behavior for post- and pre-break eras, we should not use the same coefficient (assume the same relationship) in our analysis. More simply, if the underlying dataset contains a structural break and we ignore that break, then our model would produce spurious results. Therefore we strongly suggest testing the possibility of a break in the series of interest and, if there is a break, then incorporating that break in the model.

Even with such a robust dataset, the rules of statistics play a crucial role in sound analysis.

We utilize the Core Logic Home Price Index (HPI) series to demonstrate the structural break test, Figure 9. We use a state space approach to test for a break, see Silvia and Iqbal et. al (2014) for more detail about the state space approach. We found several breaks in the Core Logic HPI series, Table 4. Assuming that the series has the same/consistent behavior over time would lead to misleading decisions.

In sum, we strongly suggest that those interested in big data applications be aware of the big data trap that hundreds of millions of observations provide complete and perfect information. Even with such a robust dataset, the rules of statistics play a crucial role in sound analysis. Ensure that the sample is a good representative of the population and test for non-stationarity, volatility clustering and structural breaks. Big data has the potential to provide more insights and improve the effectiveness of decision making if utilized correctly. Just adding more observations/respondents (or frequency), however, is not enough, and analysts need to follow proper modeling practices to reap the benefits big data has to offer.

[19] **It is worth noting that Engle (1982) introduced the ARCH method and Bollerslev (1986) provided the generalized ARCH (GARCH) approach. GARCH is the extension of the ARCH model and provides accurate results and thereby we utilize GARCH in practice. See for more detail, Bollerslev, T. (1986). Generalized Autoregressive Conditional Homoscedasticity. *Journal of Econometrics*, Vol 21 (3).**

Econometric Modeling Using Big Data: Causation vs. Correlation

Almost every time an analyst runs a regression between variables of interest, the objective is to estimate an empirical, causal relationship. Analysts strive to determine which variable is leading (causing) and which one is lagging (effect). Mostly, we utilize regression analysis to estimate the model and that estimation method provides a *statistical association*. For instance, say we want to estimate a relationship between the inflation rate and unemployment rate (Figure 10). The common practice is to run a regression analysis, which estimates a statistical association between the two variables and then the interpretation is something along the lines of a one percentage point drop in the unemployment rate would *lead* to a 0.20 percentage point increase in the inflation rate. This interpretation is one of the most simple yet often overlooked aspects of statistical analysis. The Granger causality test, not the regression analysis, estimates a statistical causality and therefore we must employ the Granger causality test for causal inference.[20] The Granger causality test estimates which variable is leading and which one is lagging (or if both depend on each other in the case of a two-way causality, see Silvia and Iqbal et. al (2014) for more detail).

The regression analysis shows a statistically significant relationship between the CPI and the unemployment rate. However, the Granger causality test indicates that there is no statistically significant causal relationship between the unemployment rate and inflation rate.

Thus, even if the data are a good snapshot of the population, we still need to utilize appropriate statistical tools (such as the Granger causality test) to estimate statistical causality. Essentially, for an accurate estimation, reliable data and appropriate econometric technique are necessary. Big data improves the data/information part of the estimation, but an appropriate model is necessary to help decision makers design effective policies.

Big data applications are a big deal for decision makers, but we urge our readers to remember the Digest example.

Concluding Remarks: Big Data Is a Big-Deal for Decision Makers

In our view, big data applications are a big deal for decision makers. On the one hand, if a decision maker believes that by using big data he/she therefore has all information and should not worry about statistical issues or causal relationships, then we urge those analysts to remember the Literary Digest example. On the other hand, if a decision maker considers big data as a means of acquiring more information and follows statistical principles along with appropriate econometric tools, then the opportunities for effective decision making will be enhanced.

[20] For more detail about the Granger causality and statistical association see Silvia and Iqbal et. at (2014). Also see Granger, C.W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, Vol 37, no3.

Appendix

Table 1: The ADF Test Results

Variable Name	Level	Growth Rate
Personal Consumption	Non-Stationary	Stationary
Personal Income	Non-Stationary	Stationary
S&P 500 Index	Non-Stationary	Stationary
Nonfarm Payrolls	Non-Stationary	Stationary

Table 2: Testing for a Statistical Relationship: Level vs. Growth Rate

Variables Name	Coefficient	t-value	R-Squared
Level Form of Personal Income and Consumption	0.8153*	519.74	0.9986
Growth Rate Form of Personal Income and Consumption	0.1348*	3.51	0.0309

*Statistically significant at 1%

Table 3: Results Based on the OLS and ARCH/GARCH Models

The Relationship Between the S&P 500 and Employment			
Estimation Method	Coefficient	t-value	R-Squared
OLS Estimates (Level Form)	0.0390*	51.89*	0.9986
OLS Estimates (MoM Form)	3.119	2.81*	0.02
ARCH/GARCH Estimates (MoM Form)	0.2567	0.26	N/A

*Statistically significant at 1%

Table 4: Evidence of Structural Breaks

Identifying a Structural Break Using the State-Space Approach*		
CoreLogic Home Price Index (YoY)		
Break Date	Type of Break	Coefficient
Dec-89	Level Shift	-0.79
Mar-11	Level Shift	-0.78
Mar-10	Level Shift	0.66
Mar-12	Level Shift	0.55
Mar-07	Level Shift	-0.54

*All Breaks Are Statistically Significant at 1%

Source: Wells Fargo Securities

Wells Fargo Securities Economics Group

Diane Schumaker-Krieg	Global Head of Research, Economics & Strategy	(704) 410-1801 (212) 214-5070	diane.schumaker@wellsfargo.com
John E. Silvia, Ph.D.	Chief Economist	(704) 410-3275	john.silvia@wellsfargo.com
Mark Vitner	Senior Economist	(704) 410-3277	mark.vitner@wellsfargo.com
Jay H. Bryson, Ph.D.	Global Economist	(704) 410-3274	jay.bryson@wellsfargo.com
Sam Bullard	Senior Economist	(704) 410-3280	sam.bullard@wellsfargo.com
Nick Bennenbroek	Currency Strategist	(212) 214-5636	nicholas.bennenbroek@wellsfargo.com
Anika R. Khan	Senior Economist	(212) 214-8543	anika.khan@wellsfargo.com
Eugenio J. Alemán, Ph.D.	Senior Economist	(704) 410-3273	eugenio.j.aleman@wellsfargo.com
Azhar Iqbal	Econometrician	(704) 410-3270	azhar.iqbal@wellsfargo.com
Tim Quinlan	Senior Economist	(704) 410-3283	tim.quinlan@wellsfargo.com
Eric Viloria, CFA	Currency Strategist	(212) 214-5637	eric.viloria@wellsfargo.com
Sarah House	Economist	(704) 410-3282	sarah.house@wellsfargo.com
Michael A. Brown	Economist	(704) 410-3278	michael.a.brown@wellsfargo.com
Jamie Feik	Economist	(704) 410-3291	jamie.feik@wellsfargo.com
Erik Nelson	Currency Strategist	(212) 214-5652	erik.f.nelson@wellsfargo.com
Misa Batcheller	Economic Analyst	(704) 410-3060	misa.n.batcheller@wellsfargo.com
Michael Pugliese	Economic Analyst	(704) 410-3156	michael.d.pugliese@wellsfargo.com
Julianne Causey	Economic Analyst	(704) 410-3281	julianne.causey@wellsfargo.com
E. Harry Pershing	Economic Analyst	(704) 410-3034	edward.h.pershing@wellsfargo.com
Hank Carmichael	Economic Analyst	(704) 410-3059	john.h.carmichael@wellsfargo.com
Donna LaFleur	Executive Assistant	(704) 410-3279	donna.lafleur@wellsfargo.com
Dawne Howes	Administrative Assistant	(704) 410-3272	dawne.howes@wellsfargo.com

Wells Fargo Securities Economics Group publications are produced by Wells Fargo Securities, LLC, a U.S. broker-dealer registered with the U.S. Securities and Exchange Commission, the Financial Industry Regulatory Authority, and the Securities Investor Protection Corp. Wells Fargo Securities, LLC, distributes these publications directly and through subsidiaries including, but not limited to, Wells Fargo & Company, Wells Fargo Bank N.A., Wells Fargo Advisors, LLC, Wells Fargo Securities International Limited, Wells Fargo Securities Asia Limited and Wells Fargo Securities (Japan) Co. Limited. Wells Fargo Securities, LLC. is registered with the Commodities Futures Trading Commission as a futures commission merchant and is a member in good standing of the National Futures Association. Wells Fargo Bank, N.A. is registered with the Commodities Futures Trading Commission as a swap dealer and is a member in good standing of the National Futures Association. Wells Fargo Securities, LLC. and Wells Fargo Bank, N.A. are generally engaged in the trading of futures and derivative products, any of which may be discussed within this publication. Wells Fargo Securities, LLC does not compensate its research analysts based on specific investment banking transactions. Wells Fargo Securities, LLC's research analysts receive compensation that is based upon and impacted by the overall profitability and revenue of the firm which includes, but is not limited to investment banking revenue. The information and opinions herein are for general information use only. Wells Fargo Securities, LLC does not guarantee their accuracy or completeness, nor does Wells Fargo Securities, LLC assume any liability for any loss that may result from the reliance by any person upon any such information or opinions. Such information and opinions are subject to change without notice, are for general information only and are not intended as an offer or solicitation with respect to the purchase or sales of any security or as personalized investment advice. Wells Fargo Securities, LLC is a separate legal entity and distinct from affiliated banks and is a wholly owned subsidiary of Wells Fargo & Company

Important Information for Non-U.S. Recipients

For recipients in the EEA, this report is distributed by Wells Fargo Securities International Limited ("WFSIL"). WFSIL is a U.K. incorporated investment firm authorized and regulated by the Financial Conduct Authority. The content of this report has been approved by WFSIL a regulated person under the Act. For purposes of the U.K. Financial Conduct Authority's rules, this report constitutes impartial investment research. WFSIL does not deal with retail clients as defined in the Markets in Financial Instruments Directive 2007. The FCA rules made under the Financial Services and Markets Act 2000 for the protection of retail clients will therefore not apply, nor will the Financial Services Compensation Scheme be available. This report is not intended for, and should not be relied upon by, retail clients. This document and any other materials accompanying this document (collectively, the "Materials") are provided for general informational purposes only.

SECURITIES: NOT FDIC-INSURED/NOT BANK-GUARANTEED/MAY LOSE VALUE