Intuit Small Business Revenue Index

White Paper

September 2012

Contents

ABOUT THE INDEX

As the leading provider of small business financial and business management tools, Intuit has a wealth of information about small business activity. In addition to our Small Business Employment Index, we now offer the Small Business Revenue Index to provide a unique view into this dynamic sector of the market. Up until now, there has been a lack of current information and understanding on small businesses and the unique impact they have on the U.S. economy.

This Revenue index is the first to provide current information on small business revenue, and the only source of monthly small business revenue data. Near real-time data is used to measure revenue (receipts) per company, which is compiled into an index starting from 100 in January 2005 for each industry group. The data comes from approximately 100,000 Intuit QuickBooks Online customers, a subset of the total QuickBooks Online user base, and is published monthly on a close to real-time basis.

INDEX PUBLICATION SCHEDULE

Generally, the Intuit Small Business Revenue Index will be released at the end of the month, reporting on revenue trends of the prior month. The data for the Revenue Index is pulled on the 21st of each month for the previous full calendar month. A few extra weeks are needed to make sure that the Index has the most up-to-date data as many businesses backfill their accounting files for the previous month.

METHODOLOGY

1.1. The data

We begin with data on revenue each month for each user of QuickBooks Online. Our goal is to measure revenue per company per month, by industry and then for all industries, for small businesses. We do not have a definition for "small" as we do in the Small Business Employment Index (fewer than 20 employees), we simply assume that a business using QuickBooks to record its activities is small.

The Revenue Index measures the level of small business monthly revenue and is reported as a series starting from 100 in January 2005, the baseline month. The data for the Revenue Index comes from approximately 100,000 companies that use QuickBooks Online. We classify each company into one of 20 industry sectors used by the IRS Statistics of Income (SOI) for its Nonfarm Sole Proprietorship Statistics.[1]

Because the average size of companies adopting the QuickBooks Online service each month changes over time, we cannot simply compute average revenue per company for each month and interpret the monthly changes as a reflection of overall economic activity for small businesses. In order to abstract from fluctuating company size, we use a statistical approach to separate the changing size of companies who adopt QuickBooks from the impact of the economic conditions on these companies. We begin by calculating monthly revenue per company separately for each monthly cohort (a set of companies adopting QuickBooks Online in a given month) and industry.

QuickBooks Online users do not enter all data immediately. Companies record their transactions into QuickBooks Online according to their own schedules, for example, weekly, monthly, or quarterly. When the monthly revenue is pulled at a particular point in time, it may not yet include all transactions occuring in that month. Additional transactions may be recorded later, or "backfilled". To adjust for the incompleteness of the most recently pulled data (which will be more complete in later, subsequent pulls), each month we extract the most recent 12 months of revenue transactions and forecast the expected amount of revenue to be backfilled for those months using a statistical model trained on past observations. Essentially we are expecting QuickBooks users to enter additional data (transactions) and we are forecasting the amounts.

Some QuickBooks Online data is discarded. The following filters are applied each month:

1. Companies with an address outside the U.S. are excluded, because we want our indexes to reflect the U.S. economy.
2. A company is excluded if all of its transactions in the last 12 months are revenue only or expenses only, to reduce bias from missing revenue or expenses.
3. A company is excluded if we are unable to determine the primary industry in which that company does business.
4. Within each month and each industry, the set of company revenues are ordered, and values more extreme than the 2nd and 98th percentiles are replaced by the values at those percentiles (this is known as "winsorizing"). This reduces the effect of extreme outliers in the data, some of which are evidently data entry errors.
5. Lastly, a transaction is excluded if it occurred in or before the month when the company

[1] See http://www.irs.gov/uac/SOI-Tax-Stats---Nonfarm-Sole-Proprietorship-Statistics

started with QuickBooks Online, to enable more accurate measurement of cohort effects. New users often enter transactions for dates before their signup (to keep their records all in one place), but some enter a longer time span of previous data (for example, some enter 6 months while others enter 2 years) than others. Thus, for consistency in building the indexes, we include only data for months after each company's signup date.

1.2. Constructing the index for a sector

Since the influence of the recession and other economic factors are likely quite different for different industry sectors, we first estimate the indexes separately for each industry sector before combining them into an overall index. In this section, we describe how the indexes for a sector are constructed.

The main steps are:

1. Perform a regression to forecast the amount of backfilled revenue
2. Perform a regression (ANOVA) to estimate the effects of month, cohort, and age on revenue. The "month effect" (coefficients on the month indicator variables) captures the influence of economic conditions.
3. Seasonally adjust the month effect and compute its trend

These are described in more detail below.

1. Backfill prediction. A dataset is prepared that contains the average revenue for each industry each month as observed 1, 2, …, 12 months after the transaction month. The total revenue recorded for a given month changes from one observation month to another because of backfill. For example, as of June 2010, we usually see that companies have more revenue recorded for in April 2010 than we saw for the same companies in May 2010. The revenue observed after 12 months is taken to be the "complete" or "true" value and is used in a regression to forecast the additional revenue to come using the following covariates: 1) observed revenue to date, 2) observation month, 3) transaction month, and 4) sector. The regression model is then applied to the main dataset to forecast what the average revenue will be 12 months after the transaction month (when firms have completed entering their transactions). We use the forecasted revenue in the subsequent steps.

2. ANOVA regression. An initial examination of the data reveals that the companies who signed up earliest for QuickBooks Online in general have more revenue and more income than later signups. In other words, the earlier signups were bigger companies than later signups. We call this a "cohort" effect, referring to incoming cohorts of users. If we simply calculate monthly revenue per company, we see revenue falling over time because the new companies entering are smaller. Thus, we cannot simply take the average revenue over all companies each month and expect this average to tell us what is going on in the economy: the average reflects both changes coming from the overall economy and changes in the average new QuickBooks Online user.

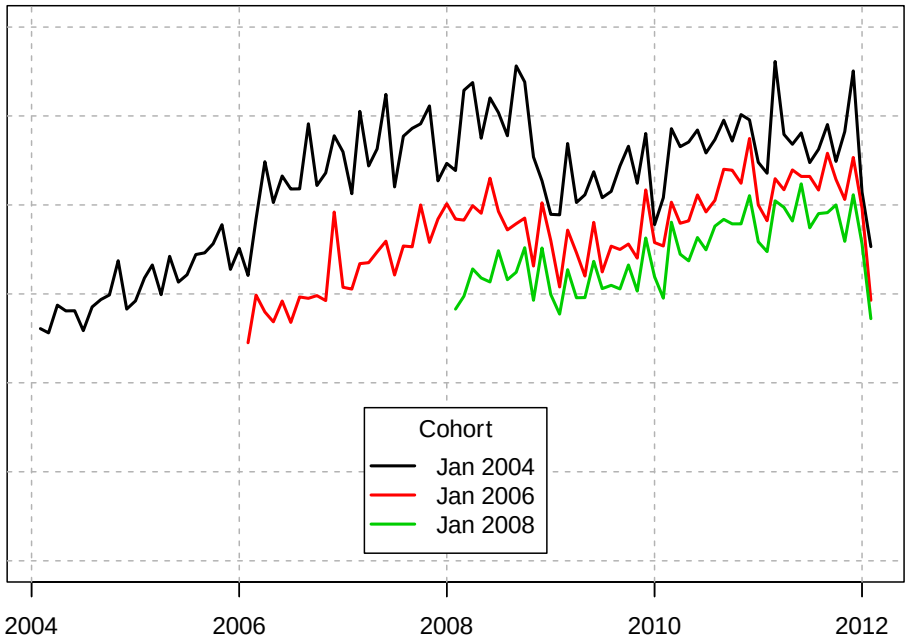

Figure 1. Average revenue for three cohorts in the "Professional, scientific, and technical services" sector.

Figure 1 above shows the monthly revenue per company for three cohorts of companies in the "Professional, scientific, and technical services" sector, one entering January 2004, another entering January 2006, and a third entering January 2008. We see that the revenue per company is highest for the 2004 cohort and lowest for the 2008 cohort. But in all three, we can see a decline in monthly revenue per company in late 2008, the time of the U.S. bank panic associated with the failure of Lehman Brothers. This demonstrates that 1) the impact of the business cycle is present in the data, but that 2) we need to take account of cohort to see the impact of the business cycle clearly.

In addition, it appears that there are systematic patterns in revenue related to the number of years companies have used QuickBooks Online. We surmise that this is related to company age. We do not actually know company age, we only know how long the company has been a user of QuickBooks Online, and so we use what we have. Revenue appears to rise over four to five years, then decline.

To isolate the influence of economic conditions on company revenue, and to separate it from the changing size of customer cohorts and changes related to company age, we do a statistical analysis known as regression. We regress revenue on categorical (dummy) variables representing month, cohort, and age. In this procedure, we assume that the average revenue at month *m* for cohort *c*, denoted by $rev_{m,c}$, consists of separate additive effects, one arising from the observation month *m*, one from the cohort start month *c* equation, and one from the company age *a*. This is represented by the following regression equation:

$$revenue_{m,c} = constant + \beta_m + \gamma_c + \delta_a + error.$$

The output of the regression procedure gives estimates for the constant term and three sets of *coefficients,* β_m, γ_c, and δ_a, computed to minimize the mean squared error weighted by the number of companies in month *m* and cohort *c*. This gives us the following sets of coefficients: β_m for the

months, revealing the business cycle from January 2004 to the present; γ_c for the cohorts, revealing average differences in the cohorts from December 2003 (this first cohort includes companies who signed up prior to December 2003) to the previous month; and δ_a for age, from one month to the maximum age in the data.

This type of regression is also called *analysis of variance* (ANOVA), because the independent variables (month, cohort, and age) are categorical rather than numeric. Using categorical variables removes the need for further assumptions on how the revenue depends on each of those variables. The idea is that the coefficients of the cohort variables will trace out the changing size of the companies in different cohorts, the coefficients of the age variables will trace out how revenue changes as the company ages, and the month coefficients will reveal the influence of economic activity in different months generally on company revenue. The month coefficients are what we are primarily interested in.

Because the regression has a constant term, we cannot estimate the absolute levels of the individual effects. For example, for the month variable, only the effects of each month *relative to some baseline month* can be estimated. This is a standard example of what is called *non-identifiability* of the coefficients, and a common method to deal with this is to impose the constraint that β_m is zero for the first month. We do not do this since we want to get an absolute level that reflects the monthly revenue with the estimated cohort and age effects removed. Our approach is to add a constant to the β_m series so that its average value in the first year — January to December 2005 — is the same as the average monthly revenue for the same period.

There is a second, more subtle source of non-identifiability: the fact that month, cohort, and age are related linearly as

$$\text{age } a = 1 + (\text{month } m) - (\text{cohort start month } c).$$

Hence, a second constraint is needed to make the coefficients identifiable. The assumption we make is that the age effects cannot be entirely up (revenue always rises with age) or down (revenue always falls with age) but must rise and then fall, as appears reasonable from examining the data. Specifically, we assume that the average age effect in the second year (13th through 24th month) is the same as the average age effect in the sixth year (61st through 72th month).

3. Seasonal adjustment. Two remaining steps transform the resulting month effects β_m into an index:

1. Seasonally adjust and compute the trend

   Unsurprisingly, the month effects have a strong seasonal character. We use the X-12-ARIMA software program from the U.S. Census Bureau to seasonally adjust the series from the previous step and obtain its (additive) trend component. We compute the trend because even the seasonally-adjusted series exhibits substantial reversion, meaning that if the figures are high one month, they tend to be low the next, and vice-versa. The trend gives us a series free of reversion.

2. Scale to a 100-based index

   Since we want to show percentage changes in the influence of economic activity rather than actual dollar amounts, which can be specific to QuickBooks Online companies, we multiply the series from the previous step by a constant so that it starts from 100 on January 2005.

1.3. Constructing the overall index

The overall index is calculated based on the following component sectors:

1. Accommodation, food services, and drinking establishments
2. Administrative and support, and waste management and remediation services
3. Agriculture, forestry, hunting, and fishing
4. Arts, entertainment, and recreation
5. Construction
6. Educational services
7. Finance and insurance
8. Health care and social assistance
9. Information
10. Manufacturing
11. Other services
12. Professional, scientific, and technical services
13. Real estate and rental and leasing
14. Retail trade
15. Transportation and warehousing
16. Wholesale trade (merchant wholesalers)

(Four other sectors used by the IRS Sole Proprietorship Statistics are currently excluded because of insufficient data or because they contain mostly not-for-profit companies.)

The steps to construct the overall index are straightforward:

1. Compute the weighted average of the sector-specific β_m series from the previous section

   The weights are selected to match the sector mix in the IRS Sole Proprietorship Statistics of Income data, as calculated by the share of the number of returns for different sectors. They are computed annually to match the IRS data for that year, or the most recent year if the IRS data for that year has not been released.

2. Adjust the series for seasonality and compute the trend, as before
3. Scale the data to a 100-based index, as before

We use a weighted average to match the IRS Sole Proprietorship population to make the index more representative of all small business and to eliminate the influence from a changing QuickBooks Online sector mix, which may occur because of factors such as marketing campaigns that target particular sectors.

The industry weighting is the only treatment of the data that we undertake to try to make the overall index representative of all small business. We have little by which we could benchmark the QuickBooks customers to all small businesses. Thus, we are simply reporting an industry-weighted average of QuickBooks customers, and how close this comes to revenues for all small business we do not know as yet.

1.4. Results for the revenue index

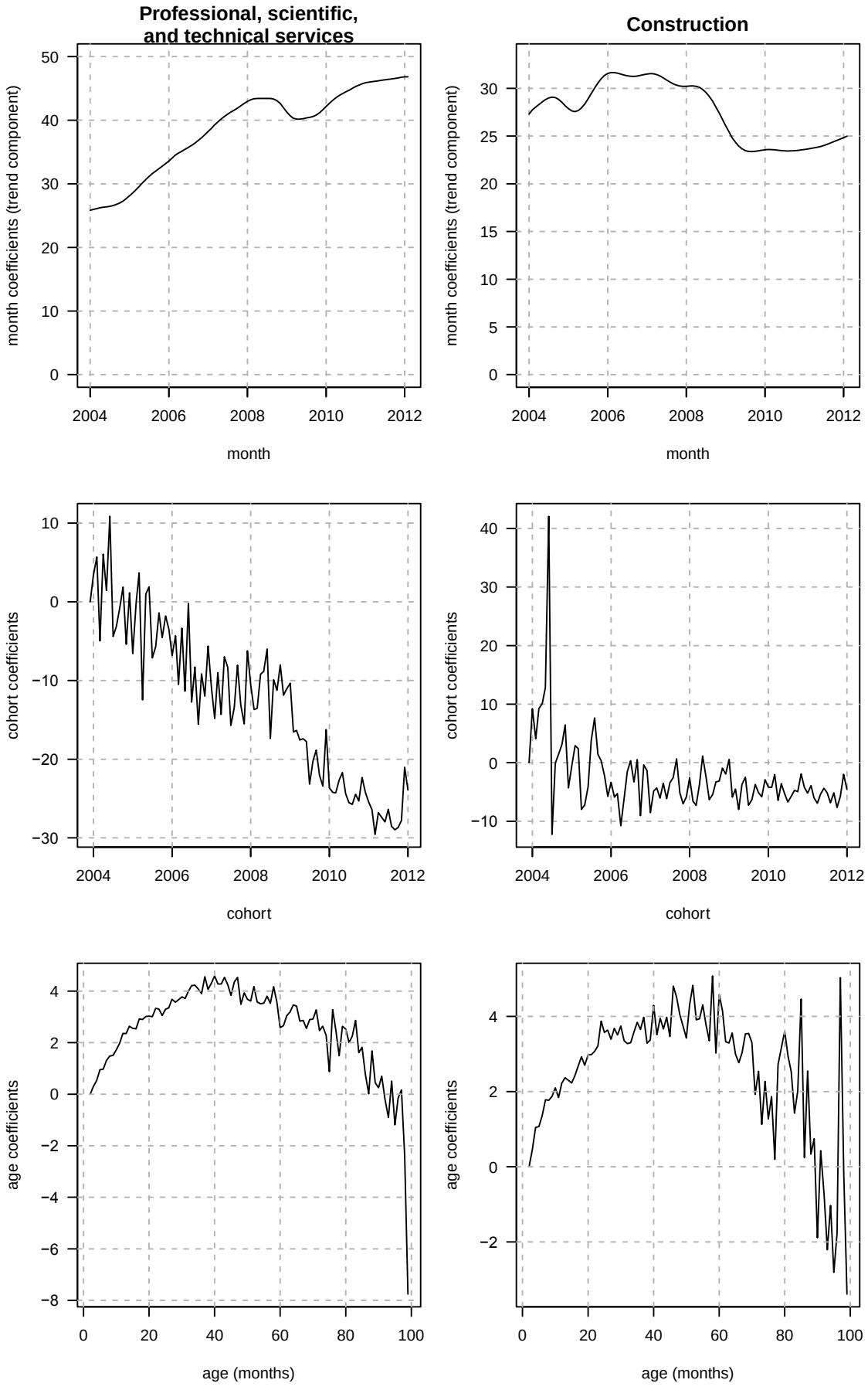

Figure 2. ANOVA regression coefficients for the two sectors: "Professional, scientific, and technical services" (left) and "Construction" (right). The month coefficients were seasonally adjusted using X-12-ARIMA and the trend components are shown. The scales on the vertical axes are not dollar amounts and are different for the two sectors.

The results show generally that the newer cohorts of QuickBooks Online users are indeed smaller than older cohorts, though this is less true in the most recent cohorts. The results also show the age effects we presume are reasonable.

More important, across the different sectors we can see two different recessions revealed in the month coefficients.

The first recession is related to the collapse in construction and durables consumption tied to the decline in the value of residential real estate, beginning in 2006. The two sectors "Construction" and "Real estate and rental and leasing" (data not shown) show this most clearly. For both of these, revenue stops rising and begins declining in 2006, with little recovery so far. Construction revenue is dealt a further blow in 2008.

The second recession coincides with the bank panic in late 2008. The two sectors "Professional, scientific, and technical services" and "Other services" (data not shown here) both show abrupt declines in revenue and income beginning around that time.

ABOUT THE ECONOMIST

The *Intuit Small Business Revenue Index* has been constructed from QuickBooks Online data by nationally-recognized financial economist, Susan Woodward.

Susan Woodward is the founder of Sand Hill Econometrics, which publishes the Sand Hill Index, an index of value for venture-capital-funded companies. She has licensed this index to Dow Jones, which began publishing it as the Dow Jones Index of Venture Capital in the fall of 2009. Sand Hill Econometrics also provides measurement of risk and performance for alternative assets to institutional investors.

Woodward has a Ph.D. in Financial Economics from UCLA. She taught finance for the first ten years of her career. She then served 10 years (1985-1995) in the government in Washington D.C., including four years as chief economist at U.S. Department of Housing and Urban Development, and four years as chief economist at the U.S. Securities and Exchange Commission. She now lives in Menlo Park, California, and works on issues in both mortgage lending and securities.

For more information, please contact:

Contacts:

Sharna Brockett
Intuit Inc.
650-944-3856
IntuitSmallBusinessIndex@intuit.com

Vanessa Piccinini
Access Communications
415-844-6296
vpiccinini@accesspr.com