A Note on the Estimation of *CAY* *

Igor Kojanov†

First version: December 2006
Current version: August 27, 2007

Abstract

The variable *cay* identified by Lettau and Ludvigson (2001) and reported on their websites contains an error in the calculation of labor income. In particular, the line "Other" from NIPA tables is double-counted. This paper documents the error and reviews the potential biases it may create. All papers that have used *cay* data downloaded from Martin Lettau's website in 2004–2007 may be affected by this bookkeeping error. As of June 2007 (after this note was sent to the authors), the mistake has been corrected.

*Special thanks to Robert Goldstein for constant encouragement. I also thank Sydney Ludvigson for her comments on the earlier draft of this note. Any errors in the paper are my own.

†Department of Finance, Carlson School of Management, University of Minnesota, koja0002@umn.edu, 612-501-3431.

1 Introduction

The *cay* variable introduced in Lettau and Ludvigson (2001) has generated significant interest due to its ability to forecast equity returns. This variable has also been widely used by other authors in the financial community, for example, as a proxy for price of risk, or as an instrumental variable in asset pricing models.[1]

Many authors obtain *cay* by directly downloading it from Martin Lettau's website without updating or recalculating. However, since late 2003 the variable published on the website contains an accounting error. In particular, a change by the Bureau of Economic Analysis in 2003 to some of their definitions has been incorrectly accounted for. In particular, this has led to a double-counting of the line item named "Other."

This mistake is purely a bookkeeping error and the extra component is relatively small. However, it has the potential to create some biases in the properties of *cay*. For example, the persistence of *cay* increases slightly, and this could possibly contribute to making forecasting regressions more spurious.

The error was eliminated and the data was updated as of June 2007.

2 The Issue

2.1 BEA Revision

In 2003, the Bureau of Economic Analysis (BEA) released a comprehensive revision of the National Income and Product Accounts (NIPA). This revision had an impact on the implementation of many data handling procedures in economics and finance research.

In particular, one of the affected time series is *cay*. On one hand, according to the note on Martin Lettau's web-site [2]:

> ** OTHER LABOR INCOME: As of Dec. 2003 "other" labor income was renamed "Employer Contributions for employee pension and insurance funds". The new "other" labor income is a catch-all residual consisting of unspecified additional labor income sources, which we do not include.

On the other hand, according to the BEA's own publications (Moulton and Seskin (2003, p. 31) and Mayerhauser, Smith, and Sullivan (2003, pp. 13, 16)), the "Other labor income" from the pre-2003 tables is stripped of one subcomponent (various fees) and the rest is renamed "Employer contributions for employee pension and insurance funds", thus making the line "Other labor income" in NIPA tables obsolete. Note that this change was carried back to 1948. Thus, the entire time series for labor income and *cay* are affected.

In the authors' attempt to account for this change in BEA data, there are currently two errors made in the determination of *cay*. First, the site incorrectly assumes that "Other labor income" is kept in BEA data. Second, contrary to what the website states, it still

[1] Note, however, a series of papers casting doubt on *cay*'s success on various grounds: Brennan and Xia (2005), Goyal and Welch (2007), Rudd and Whelan (2006), Hahn and Lee (2006), Zhu (2006).

[2] http://pages.stern.nyu.edu/~mlettau/. The discussed version of the data was available on the web-site as late as April 21, 2007.

	Mean	St.dev.	Autocorr	ρ
y^{LL}	9.4242	0.3803	0.9860	0.9999
y^*	9.3825	0.3603	0.9861	0.9999
Δy^{LL}	0.0060	0.0089	0.0403	0.9857
Δy^*	0.0057	0.0090	0.0518	0.9857
cay^{LL}	0.0000	0.0121	0.8412	0.9626
cay^*	0.0000	0.0132	0.8528	0.9626

Summary statistics for labor income y, growth rate of labor income Δy and *cay*. Variables downloaded from Martin Lettau's website are denoted with "LL", corrected variables – with an asterisk (this convention is used throughout this paper). ρ is the correlation between corresponding variables.

Table 1: Summary statistics for the affected variables.

includes the line "Other" (which now has a completely different meaning in the tables distributed by BEA) in the calculation of labor income y and *cay*.

We note that the new "Other" is no longer "Other labor income". (This item does not exist in the revised tables). Instead, it is now part of "Personal current transfer receipts" (which are included in Lettau-Ludvigson's definition of labor income). Thus, the website essentially counts the "Other" component of "Personal current transfer receipts" *twice*. [3]

2.2 Effect of the Error

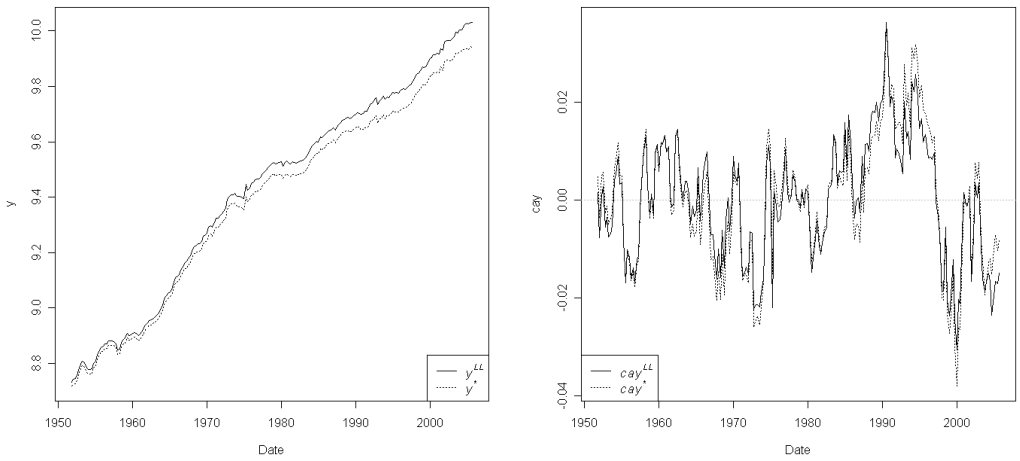
How large and important is the difference? At first, it seems to be less than trivial: on average the difference between the correct and incorrect values amounts to -0.4% of log-labor income y with correlation between the values 0.999 (Table 1 and Figure 1, Panel A). [4]

A more important question is: how is the *cay* variable affected? The problem is that *cay* data is widely used by the finance community and any errors in its calculation may have an effect on some published and yet to be published papers. *Cay* is affected a bit more than labor income (Table 1 and Figure 1, Panel B): correlation between the two versions of *cay* is 0.96 and the correct version cay^* is a bit more persistent (autocorrelation of 0.853 versus 0.841), which may make forecasting regressions more spurious. Nevertheless, this difference also does not seem to be very large.

For completeness, I also redid other tests of *cay* properties used by Lettau and Ludvigson in their arguments: the Augmented Dickey-Fuller test of cointegration of c, a and y and

[3] If I follow the website's definition of labor income *and count "Other" twice* I get an exact replica of *cay* value from the website.

[4] In all calculations below I use the sample period 1951:Q4-2005:Q4 (if I use a shortened data range, 1951:Q4-2002:Q4, to avoid the most recent minor BEA revisions unrelated to the issue at hand and not available to Lettau and Ludvigson when they last updated their data, the results do not change substantially). In constructing my variables I used website definitions of all quantities, BEA NIPA data downloaded on September 5, 2006, and for the equity excess return I used the CRSP value-weighted return over 3-months T-bill rate.



Panel A Panel B

Figure 1: Two versions of log labor income y (Panel A) and *cay* (Panel B): correct (dotted line) and incorrect (solid line).

the return forecasting regressions (in-sample and out-of-sample).

Cointegration tests (Table 2) show that the correction makes the test less likely to reject the hypothesis of no cointegration between c, a and y. Notice that with updated date range the hypothesis of no cointegration cannot be rejected at a 5% level for either of the variables (downloaded and corrected), contrary to the original result in Lettau and Ludvigson (2001). Of course, this is just one of many applicable cointegration tests; other tests may show different results. The full investigation of the cointegration issue is well beyond the scope of this note.

More striking is that the return predictability results (both in-sample and out-of-sample, Tables 3 and 4) are improved significantly (as measured by R^2 in the in-sample case and the ratio of root mean-square errors in the out-of-sample case) especially at long horizons. However, the improved return predictability may be due to the increased persistence of *cay* (this problem is discussed in Ferson, Sarkissian, and Simin (2003): corrected *cay*'s autocorrelation gets closer to the "critical" boundary of 0.9 mentioned in that paper).

3 Conclusion

The impact of the bookkeeping error in the calculation of labor income and *cay* may be small; however, it is possible that in some cases it may slightly change results. As documented here, in some cases the exclusion of "Other" weakens conclusions about properties of *cay*.

Since Lettau and Ludvigson's contribution of *cay* to the economic literature has been influential, a lot of researchers have used its value downloaded directly from Martin Lettau's

Lag	1	2	3	4
t-stat^{LL}	-3.648	-3.301	-3.233	-2.877
t-stat*	-3.613	-3.400	-3.379	-3.024

Augmented Dickey-Fuller test for the fitted residuals from the cointegrating regression of consumption on labor income and wealth. "Lag" is the number of lags of first differences used in the Dickey-Fuller test regression. For both cases, the sequential data-dependent lag selection procedure implies the significant lag length is one. Critical values are: -3.80 (5%), -3.52 (10%).

Table 2: Residual-based tests for cointegration of c, a and y.

Horizon	1	4	8	12	16	20	24
cay^{LL}	1.539	4.993	8.475	11.052	11.580	12.493	13.660
NW t-stat	3.112	3.178	3.640	3.725	3.822	4.332	4.330
R^2	0.051	0.130	0.211	0.274	0.271	0.253	0.252
cay^{*}	1.549	5.120	8.692	11.399	12.245	13.431	14.690
NW t-stat	3.488	3.591	4.513	4.974	5.846	6.758	5.720
R^2	0.062	0.168	0.279	0.369	0.385	0.371	0.364

The table reports forecasting regression coefficient on *cay*, Newey-West corrected t-statistics and R^2 for both variables. "Horizon" is the period over which the excess return is calculated.

Table 3: In-sample excess return forecasts.

Horizon	1	4	8	12	16	20	24
cay^{LL}	1.000	1.008	0.989	0.954	0.962	0.954	0.991
cay^{*}	1.001	0.998	0.977	0.936	0.940	0.943	0.986

The table reports the ratio of the out-of-sample root mean-square errors (RMSE) from the *cay*-based forecasting regression to the root mean-square errors for "naive" forecast using the sample mean excess return (lower value of RMSE ratio means better performance of *cay*-based forecast over the naive mean forecast). "Horizon" is the period over which the excess return is calculated.

Table 4: Out-of-sample excess return forecasts.

web-site. It is impossible to compile a list of all the papers that used the data series downloaded in 2004–2007.[5] For a number of them the effect of this miscalculation may be unimportant, but for some it may have consequences, and should be checked on a case by case basis.

References

Brennan, Michael J., and Yihong Xia, 2005, *tay*'s as good as *cay*, *Finance Research Letters* 2, 1–14.

Ferson, Wayne E., Sergei Sarkissian, and Timothy T. Simin, 2003, Spurious Regressions in Financial Economics?, *Journal of Finance* 58, 1393–1413.

Goyal, Amit, and Ivo Welch, 2007, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, forthcoming in Review of Financial Studies.

Hahn, Jaehoon, and Hangyong Lee, 2006, Interpreting the Predictive Power of the Consumption-Wealth Ratio, *Journal of Empirical Finance* 13, 183–202.

Lettau, Martin, and Sydney Ludvigson, 2001, Consumption, Aggregate Wealth, and Expected Stock Returns, *Journal of Finance* 56, 815–849.

Mayerhauser, Nicole, Shelly Smith, and David F. Sullivan, 2003, Preview of the 2003 Comprehensive Revision of the National Income and Product Accounts: New and Redesigned Tables, *Survey of Current Business* 83, 7–31.

Moulton, Brent R., and Eugene P. Seskin, 2003, Preview of the 2003 Comprehensive Revision of the National Income and Product Accounts: Changes in Definitions and Classifications, *Survey of Current Business* 83, 17–34.

Rudd, Jeremy, and Karl Whelan, 2006, Empirical Proxies for the Consumption-Wealth Ratio, *Review of Economic Dynamics* 9, 34–51.

Zhu, John Qi, 2006, Cointegration of Consumption, Wealth, and Income: Evidence from Micro Data, Working Paper.

[5] Only works that used the data downloaded in 2004–2007 may be affected, earlier data is spared of the error. Since the error was eliminated as of June 2007, all the data downloaded after that are also not affected.