Relationship Between Data Quality and Collection Date in the Consumer Price Index Housing Survey

Shawn Jacobson
Shawn Jacobson, Bureau of Labor Statistics, Rm. 3655, 2 Mass. Ave. NE, Washington, DC 20212
Jacobson_S@BLS.GOV

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

KEY WORDS: Stability Bias, Seam Effect, Random Group, Exploratory Data Analysis

1. Introduction

The Consumer Price Index (CPI) Housing Survey collects Rent and other information from about 40,000 renters. This rental information is used to compute Price Indexes for residential rent (Rent) and is also used (along with information from homeowners) to compute price indexes for owner's equivalent rent (REQ). Together, these two indexes make up about 28% of the CPI.

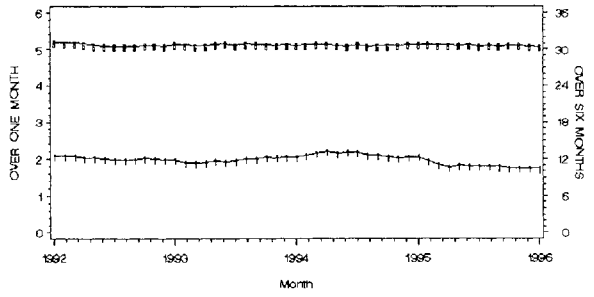
CPI staff attempt collection of rent data for rental housing units every six months. Each respondent for a rented housing unit (unit) is asked what the rent is in the current month (month T) and what the rent was in the previous month (month T-1); these values are used to compute a one-month rent price relative. The month T rent, along with the current month rent from the previous interview (month T-6) is used to compute a six-month rent price relative.

In the past, Rent and REQ indexes were computed using a composite of one and six month price relatives. However, one-month price relatives are now believed to underestimate inflation and thus add bias to price indexes for Rent and REQ (Jacobson 1994). Thus, the composite index was replaced by a chained six-month index formula in January 1995 even though the chained six-month index is seen as being less timely (takes longer to reflect changes in the rental market) than is the composite index formula.

The bias in the one-month price relative arises from the fact that some one month rent changes are not reported. Plot 1 illustrates this problem by showing twelve-month moving averages of the percentage of one and six month rent changes reported for housing units that had the same tenant over the past six months. It can be seen that 30% of housing units have rent changes over the past six months, thus, 5% of housing units should have a rent change during the previous month. However, only 2% of these housing units, report one-month rent changes instead of the 5%, that would be expected given the percentage of reported six-month rent changes.

In an attempt to make the one-month price relative fit for use, the author performed exploratory data analysis (EDA) to look for relationships between the rate of one-month rent change reporting and other characteristics of data collection. The author found that the proportion of housing units reporting one-month rent changes might be related to the date of housing data collection; in fact, the rate at which one-month rent changes were reported dropped dramatically when data collection occurred after the 20th of the month. For this reason, the author used the 20th of the month as the cutoff for early data collection. Plot 2 shows twelve-month moving averages of the percentage of reported one-month rent change for the early and late collection periods. The author also found a relatively small but persistent relationship between the proportion of reported six-month rent changes and the data collection date; Plot 3 shows twelve-month moving averages of reported six-month rent change. Since the rent paid for a housing unit during a month should be independent of the date, within the month, when data was collected, the validity of some reported rent changes between months T-6 and T are called into question.

PLOT 1
%OF OBSERVATIONS WITH 1 AND 6 MONTH
RENT CHANGE--12 MONTH MOVING AVERAGE
1-1-1=ONE MONTH 6-6-6=SIX MONTH



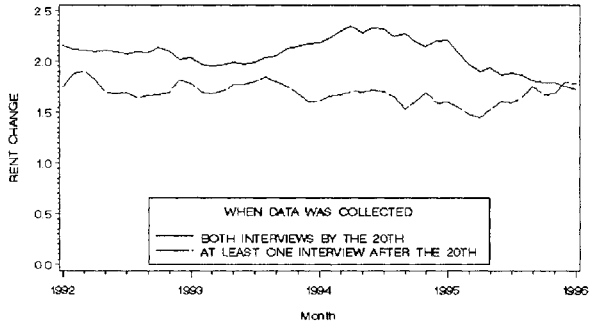
The rent values studied here are the "normalized rents" of sampled units. Normalized rents can be made up of three components: rent paid by the tenant (cash), a rent subsidy paid by a third party (subsidy), and the value of service provided by the tenant in lieu of rent (service). These normalized rents are further modified

using auxiliary data to produce rent values used to compute Rent and REQ price indexes.

These modifications are required so that services included in the CPI rent estimates will stay constant over time. For instance, if the rent is $400 in both March and September of 1998 but the landlord stops paying for electricity between these two months, then the $400 no longer pays for as much as it once did. Thus, inflation has occurred for the tenant, and the CPI needs to reflect this inflation in its price index estimates. These adjustments are important enough for the CPI to include questions used to verify changes when the value of auxiliary variables (such as who pays for electricity) change over time. So that a general study of data quality could be made, known problems with collecting data on auxiliary information were added to this study.

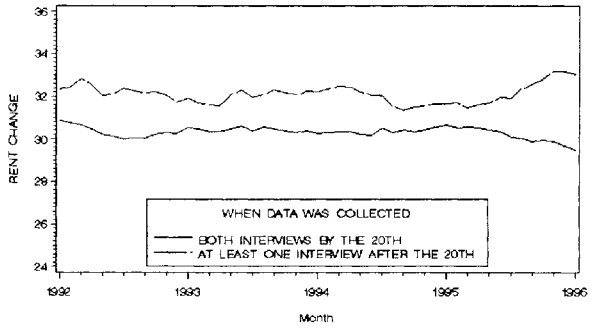
PLOT 2

%OBSERVATIONS WITH ONE MONTH RENT CHANGE
TWELVE MONTH MOVING AVERAGE



PLOT 3

%OBSERVATIONS WITH SIX MONTH RENT CHANGE
TWELVE MONTH MOVING AVERAGE



This paper gives the findings of this study. Section 2 describes the data collection problems studied by the author, and Section 3 describes the data used in the study. Section 4 gives results from the EDA phase of this research; and Section 5 gives results from tests on more recent data to see if these results hold over time. Finally, Section 6 gives recommendations for future research.

2. Potential data collection problems

The author studied a variety of data collection problems. One such problem, the failure to capture rent changes that occurred during the previous month may be the result of stability bias, the tendency to give the same answer to similar questions.

Four studied data collection problems might be caused by seam effects. Seam effects occur when the respondent forgets information given to the interviewer in the previous interview; they are a problem when the current and previous interviews are separated by a long time period such as the six months between housing interviews. The studied problems caused by seam effects are:

- Erroneous reports of rent change between month T-6 and month T.
- Answers to validation questions inconsistent with auxiliary variable reports in months T-6 and T.
- Erroneous reports, or omissions of subsidy or service adjustments.
- Erroneous reports of change in the amount of subsidy or service adjustments.

In addition, four data collection problems may be caused by the complexity of the interview. These problems are:

- Reported rent may be rounded.
- Required validation question may by omitted.
- Unnecessary verification question may be asked.
- Subsidy and service rent components may be improperly included in the cash rent component.

The last problem was suggested by Schemer (1993) who found that mothers erroneously included alimony payments when asked how much child support they had received. The problem posed by the various rent components may be similar to the problem of distinguishing between child support and alimony. Thus, respondents might include subsidy and service payments in the cash component of rent thus double counting subsidies and services. The author found that, rent changes that could be ascribed to this type of error were extremely rare, therefore the author dropped research into this data collection problem.

3. Data used for the study

This study used data collected for sampled units between July 1991 and December 1998. Each observation represents the reported rent history starting in month T-6 and ending in month T. Thus, an observation includes reported data from two interviews. The previous interview provides reports of rent and auxiliary data for month T-6, and the current interview provides this data for months T-1 and T. The collection

period for an observation is the year and month of the current interview. For example, an observation that had its current interview in July 1994 and its previous interview in January 1994 would have a collection period of 9407 (the seventh month in 1994).

Observations may overlap; that is, an interview can be the current interview for one observation and the previous interview for another. For example, if a unit is interviewed in December 1993, June 1994, and December 1994 the June interview could be the current interview in an observation for June 1994 and the previous interview for an observation for December 1994.

In order for a pair of interviews to be an observation, several conditions had to apply. The response for a housing unit had to be usable in months T-6 and T, and the reported normalized rent could not be imputed. Finally, to avoid the situation where rent changed because of a change in tenant, an observation could only be used in the study if the same tenant occupied the unit during both interviews.

The author measured the extent of stability bias by computing the percent of observations where reported month T-1 rent differs from reported month T rent. A relatively small value of this percentage corresponds to a relatively large amount of stability bias.

The author assumes that erroneous current month rent reports will cause the month T rent to be different from the rent for month T-6. Thus the prevalence of erroneous reported six-month rent changes is measured by the percent of observations where the reported month T-6 and month T rent are different. This problem is also measured by the percent of reported rent decreases (most reported rent decreases are believed to be erroneous) from month T-6 to month T. The number of mistakes made in reporting or omitting subsidy or service adjustments is measured by finding the percent of observations where the existence of a subsidy or service adjustment changed from month T-6 to T. Finally, the number of erroneous changes in these adjustments is measured by comparing the amounts of these adjustments in month T-6 to the month T amounts. The number of inconsistent responses to validation questions is found by looking at cases where auxiliary variables have changed from month T-6 to T but where this change was denied in the validation question. For all of these percentages, relatively high values correspond to relatively large seam effect error problem.

Reported rents were considered to be rounded if they were divisible by $25, so the prevalence of rent rounding was measured by looking at the percent of month T rents divisible by this amount. If an auxiliary variable had changed between month T-6 and month T, a validation question should have been asked and the percent of observations where such a question was not asked was used to measure problems with the omission of this question. Finally, if there was no change in an auxiliary variable, no validation question was needed and the percent of observations with an unnecessary validation question was used to measure the extent of this problem. Relatively large values for these percentages correspond to relatively large complexity problems.

Although sampled units have varying weights that are used in the computation of Rent indexes, the author used unweighted values in this analysis. The author's reasoning here is that the object of study is the data collection process itself rather than actual rent inflation. Thus, the author wanted each observation to be of equal importance for this study.

Observations with their current interview prior to January of 1997 (collection period less than 9701) were used in the EDA discussed in Section 4. The other observations (collection period greater than 9612) were used in tests to confirm the findings. The results of these tests are given in section 5. As of January 1998, Buffalo NY, New Orleans LA, and non-metropolitan urban areas in the Northeast Census Region were dropped as index areas. As a result, housing units in these areas were dropped from the sample. The author dropped observations from 1997 in these Index Areas before performing confirmatory data analysis in order to maintain consistency within the data used to confirm the findings from earlier data collection. This consistancy was required for purposes of variance computation.

4. Results of EDA

Table 1 gives percentages of the characteristics under study by when data was collected in month T-6 and month T. The author found that data collection was generally better when both month T-6 and month T data were collected before the 21st **(numbers in bold)** than it was when data was collected after the 20th for at least one interview (*numbers in italics*). For example, 8% of observations with consistently early data collection reported rent decreases while all other classes had about 10% of their observations report rent decreases.

Surprisingly, this was even true of the percentage of one month rent changes. That is, observations with late month T-6 and early month T collection (with 1.8% of observations with one-month rent changes) behave more like observations with late month T collection (1.7% & 1.5%) than like observations with early data collection for both interviews (2.1%). This

is surprising because the reported rent for month T-1 should be independent of data collection in month T-6. This may indicate that late data collection is not the cause of data collection problems; instead, it might be associated with them through another factor such as the level of cooperation received from respondents.

Because consistent early data collection seems associated with good data quality, the author divided observations into a "All Early" group (all data collection before the 21st) and a "Some late" group (some data collection after the 20th). The "1992-96" rows of Table 2 give comparisons of studied percentages for the "all early" and "some late" groups.

The author found that observations in the "some late" group were 6% more likely to report six-month rent changes than were other observations. For some reason, the high level of six-month rent changes in the "Some late" group was entirely due to a high percentage of rent decreases. One would conclude that all erroneous reports of rent change from months T-6 to T result in a reported rent decrease or that some legitimate six-month rent increases are not captured by late data collection. Observations with late data collection were 10% more likely to result in rounded rents than were other observations. The percent of changes in whether or not the rent is adjusted for a subsidy or service is also higher for the "some late" than for the "all early" group.

By contrast, the percent of reported changes to the amounts of subsidy and service adjustments is 8% lower for observations in the "some late" group than for other observations. That is, the observations in the "all early" group were more likely to report changes to subsidy and service amount than observations with late data collection. It is unclear why the different rent components act differently. One explanation may be that chronically late respondents don't report these adjustments at all. The analysis in this study would not be able to measure such consistent reporting errors.

For all percentages involving auxiliary variable validation, the "some late" group had higher values than the "all early" group. This is especially true of the failure to ask these questions; where the "some late" group (1.7%) was 50% more likely to miss this question than other observations (1.1%). Also, inconsistent validation answers were 25% more likely when data collection came after the 21st of the month.

5. Confirmation of findings

To see if the findings from data collected between 1992 and 1996 continued over time, the author looked at housing sample data from 1997 and 1998. The findings are given in the "1997-98" rows of Table 2. Standard errors were computed using a stratified random groups methodology similar to that described in Leaver and Valliant (1995) save that the methodology was programmed is SAS rather than in VPLX. Standard errors are given in the "s.e. 1997-98" rows of Table 2.

Generally, differences in reporting rent persisted across time and were statistically significant. The difference between "some late" and other observations in reporting one-month rent change did not change after 1996, but the percent of these changes dropped in both groups. This might be related to the dropping of the one-month price relative from price index calculation. The six-month rent change, six-month rent decrease, and month T rent rounding percents also did not change after 1996.

However, differences in other percentages between the two groups were not statistically significant after 1996, this despite the fact that the difference in percent subsidy or service amount change between the two groups increased. The difference between the two groups in percent of units changing subsidy or service status virtually disappeared.

Differences between the two groups in regard to the validation of changes in auxiliary variables disappeared after 1996. This was because data collection got better (especially for late respondents) after 1996. For instance the percent of late observations with inconsistent responses to validation questions declined by a third from 3.7% before 1997 to 2.37% after 1996. Apparently, some of the problems that late respondents have with answering questions about auxiliary variables have already been solved.

6. Conclusions and further study

The findings of this study show that the accuracy of rent reporting may decrease as rent data collection moves later into the month. If the distribution of collection dates does not change, about 25% of usable responses will be affected by late reporting. Thus, if 4,000 usable responses are gathered every month, about 1,000 responses would be effected by late reporting resulting in the failure to capture 4 one-month rent changes, about 25 erroneous rent changes, 15 erroneous rent decreases, and 35 instances when the rent is rounded. Further research is needed to see how large an affect these additional data errors have on estimates and standard errors of price index estimates.

If the effects are large, it may be advisable to consider not collecting housing data after the 20th of the month. In any case, it would be good to see if late data reporting and collection errors have a common cause (say problem respondents).

Table 1
Percent of Observations with Certain Characteristics
by Date of Interview in Months T-6 and T

Characteristic	Collection Date			
	Month T-6	Month T		
		Early	Late	All
Number of Observations	Early	136,885	17,069	153,954
	Late	16,002	5,235	21,237
	Total	152,887	22,304	175,191
Month T-1 rent does not equal month T rent	Early	**2.1**	*1.7*	2.0
	Late	*1.8*	*1.5*	1.7
	Total	2.0	1.7	2.0
Month T-6 rent does not equal month T rent	Early	**30.4**	*33.0*	30.7
	Late	*31.4*	*32.3*	31.6
	Total	30.5	32.8	30.8
Month T-6 rent is greater than month T rent	Early	**7.9**	*9.7*	8.1
	Late	*9.9*	*10.5*	10.0
	Total	8.2	9.9	8.4
Month T rent is divisible by $25	Early	**51.0**	*55.7*	51.6
	Late	*57.0*	*59.0*	57.5
	Total	51.7	56.5	52.3
Change in report of subsidy or service adjustment	Early	**2.8**	*3.3*	2.8
	Late	*3.1*	*3.3*	3.3
	Total	2.8	3.2	2.9
Change in amount of continuing subsidy or service adjustment	Early	**6.7**	*6.4*	6.7
	Late	*6.0*	*6.4*	6.1
	Total	6.6	6.4	6.6
Validation question is not asked	Early	**1.1**	*1.7*	1.2
	Late	*1.6*	*1.4*	1.6
	Total	1.2	1.6	1.2
Validation response is inconsistent with utility data	Early	**2.9**	*3.7*	3.0
	Late	*3.8*	*3.5*	3.7
	Total	3.0	3.6	3.1
Unnecessary validation question	Early	**1.8**	*2.1*	1.8
	Late	*1.9*	*2.4*	2.4
	Total	1.8	2.2	1.8

Table 2
Comparison of Characteristic Percentages
by Collection Date Class and Collection Period

Characteristic	Collection	Collection Date Within the Collection Period		
	Periods	All Early	Some Late	Difference & s.e. Difference
Number of Observations	1992-96	136,885	38,306	
	1997-98	45,469	17,301	
Month T-1 rent does not equal month T rent	1992-96	2.1%	1.7%	
	1997-98	1.67%	1.32%	-0.35%
	s.e. 1997-98	0.12%	0.12%	0.15%
Month T-6 rent does not equal month T rent	1992-96	30.4%	32.2%	
	1997-98	31.39%	33.97%	2.58%
	s.e. 1997-98	0.72%	0.93%	0.87%
Month T-6 rent is greater than month T rent	1992-96	7.9%	9.9%	
	1997-98	8.12%	9.58%	1.47%
	s.e. 1997-98	0.21%	0.40%	0.33%
Month T rent is divisible by $25	1992-96	51.0%	55.7%	
	1997-98	51.20%	54.48%	3.27%
	s.e. 1997-98	0.92%	1.52%	1.32%
Change in report of subsidy or service adjustment	1992-96	2.8%	3.2%	
	1997-98	2.98%	3.05%	0.07%
	s.e. 1997-98	0.16%	0.21%	0.22%
Change in amount of continuing Subsidy or service adjustment	1992-96	6.7%	6.2%	
	1997-98	6.74%	5.97%	-0.77%
	s.e. 1997-98	0.37%	0.41%	0.46%
Validation question is not asked	1992-96	1.1%	1.6%	
	1997-98	0.97%	1.17%	0.21%
	s.e. 1997-98	0.05%	0.10%	0.10%
Validation response is inconsistant with utility data	1992-96	2.9%	3.7%	
	1997-98	2.30%	2.37%	0.07%
	s.e. 1997-98	0.15%	0.20%	0.19%
Unnecessary validation question	1992-96	1.8%	2.0%	
	1997-98	1.21%	1.24%	0.02%
	s.e. 1997-98	0.09%	0.12%	0.15%

This study shows that it may be good to not just think in terms of respondents and nonrespondents when looking at response issues. It may be better to think in terms of good respondents, marginal respondents and nonrespondents. This new way of thinking about the response issue would call into question heroic attempts to increase the response rate at all costs. Instead, we would ask at what point effort to get information stops being useful for the production of good estimates of inflation.

References

Bureau of Labor Statistics, *BLS Handbook of Methods* (1996), Washington. DC: U.S Government Printing Office, 185 and 217-225.

Jacobson, S (1994) "Evaluation of Alternative Rent Estimators for the Consumer Price Index", *Proceedings of the Government Statistics Section*, American Statistical Association, pp 134-139.

Leaver, S. G. and Valliant, R. L. (1995) "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods*. Wiley & Sons, Inc., 1995.

Schaeffer, Nora Cate, (1993), Errors of Experience: Response Errors in Reports About Child Support and their Implications for Questionnaire Design, in *Autobiographical Memory and the Validity of Retrospective Reports*, Norbert Schwartz & Seymour Sudman Editors, Springer-Verlog, New York, New York.

Wolter, Kirk M (1985) *Introduction to Variance Estimation*, New York NY, Springer-Verlag.