

# U.S. Population: A Second Analysis

Again, my target for the day is the data describing the growth of the population of the United States. I want to derive a summary of the growth rate. I want to get an overview of the processes that generate it. This time I am going to construct a distinctly different analysis, as compared to the first.

How is it that I can construct two seriously different analyses of one set of data? I can do that because there is something detached about a text or a course on data analysis — it is detached from the actual research. Limited to these numbers, without recourse (for the moment) to other data (data on birth rates, death rates, life expectancy, age distribution, occupational statistics, immigration rates, emigration rates, ...) I can follow divergent thoughts from one set of data without the commitment and the resources by which serious research would formulate hypotheses and choose among them. Once again, here are the data, Figure 1,

---

Census Date	Resident Population	
Conterminous U.S. (Note 1)		
1790 Aug-02	3,929,214	
1800 Aug-04	5,308,483	
1810 Aug-06	7,239,881	
1820 Aug-07	9,638,453	
1830 Jun-01	12,866,020	
1840 Jun-01	17,069,453	
1850 Jun-01	23,191,876	
1860 Jun-01	31,443,321	
1870 Jun-01	39,818,449	Note 2
1880 Jun-01	50,155,783	
1890 Jun-01	62,947,714	
1900 Jun-01	75,994,575	
1910 Apr-15	91,972,266	
United States		
1920 Jan-01	105,710,620	
1930 Apr-01	122,755,046	
1940 Apr-01	131,669,275	
1950 Apr-01	150,697,361	
1960 Apr-01	178,464,236	
1950 Apr-01	151,325,798	
1960 Apr-01	179,823,175	
1970 Apr-01	203,302,231	Note 3
1980 Apr-01	226,545,805	
1990 Apr-01	248,709,873	

Figure 1  
United States Population: 1790 to 1990

Note 1: Excludes Alaska and Hawaii. Note 2: Revised to include adjustments for under numeration in southern states; unrevised number is 38558371. Note 3: Figures corrected after 1970 final reports were issued. From *Statistical Abstract of the United States, 1992, No. 1*. Original: U.S. Bureau of the Census, U.S. Census of Population: 1920 to 1990, vol. 1; and other reports.

---

This time I am thinking simply that populations grow exponentially. Maybe, so let me test that. To test that I follow the usual procedure: I set up a graph such that if the hypothesis is correct then the graph will be linear and the residuals will be noise. For exponential growth that means computing the logarithms, fitting a line to the logarithms of population, and looking at the residuals. Because I expect growth rates on the order of 2 to 3 percent and residuals on the same scale, I will use logarithms base e. (For 1950 and 1960, the two estimates, with and without Alaska and Hawaii, differ by about half of one percent, a small but non-trivial difference. For the moment, I will use their mean.)

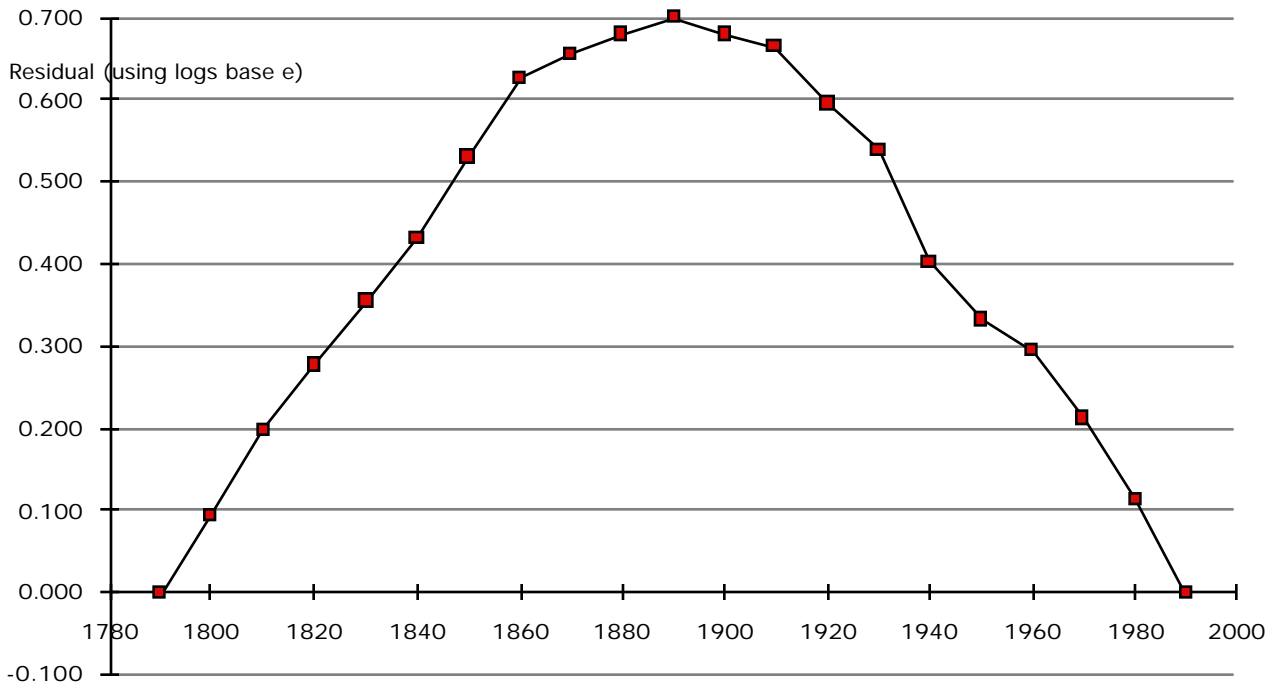
For calculation purposes I will count 1790 as year "0". That allows me to use the log of the population in 1790 as a first estimate of the intercept. Then for an estimate of the slope I compute a "rise" from the difference between the logarithm of the 1990 population and the logarithm of the 1790 and I compute a "run" of 200 years, estimating a slope of

Intercept	15.184
slope	0.02073924

Year	Years after 1790	ln(Pop)	Expected	Residual	Sqd Residual	
1790	0	3,929,214	15.184	15.184	0.000	0.000
1800	10	5,308,483	15.485	15.391	0.093	0.009
1810	20	7,239,881	15.795	15.599	0.196	0.039
1820	30	9,638,453	16.081	15.806	0.275	0.076
1830	40	12,866,020	16.370	16.014	0.357	0.127
1840	50	17,069,453	16.653	16.221	0.432	0.186
1850	60	23,191,876	16.959	16.428	0.531	0.282
1860	70	31,443,321	17.264	16.636	0.628	0.394
1870	80	39,818,449	17.500	16.843	0.657	0.431
1880	90	50,155,783	17.731	17.051	0.680	0.463
1890	100	62,947,714	17.958	17.258	0.700	0.490

1900	110	75,994,575	18.146	17.465	0.681	0.464
1910	120	91,972,266	18.337	17.673	0.664	0.441
1920	130	105,710,620	18.476	17.880	0.596	0.355
1930	140	122,755,046	18.626	18.087	0.538	0.290
1940	150	131,669,275	18.696	18.295	0.401	0.161
1950	160	151,011,580	18.833	18.502	0.331	0.109
1960	170	179,143,706	19.004	18.710	0.294	0.086
1970	180	203,302,231	19.130	18.917	0.213	0.045
1980	190	226,545,805	19.238	19.124	0.114	0.013
1990	200	248,709,873	19.332	19.332	0.000	0.000

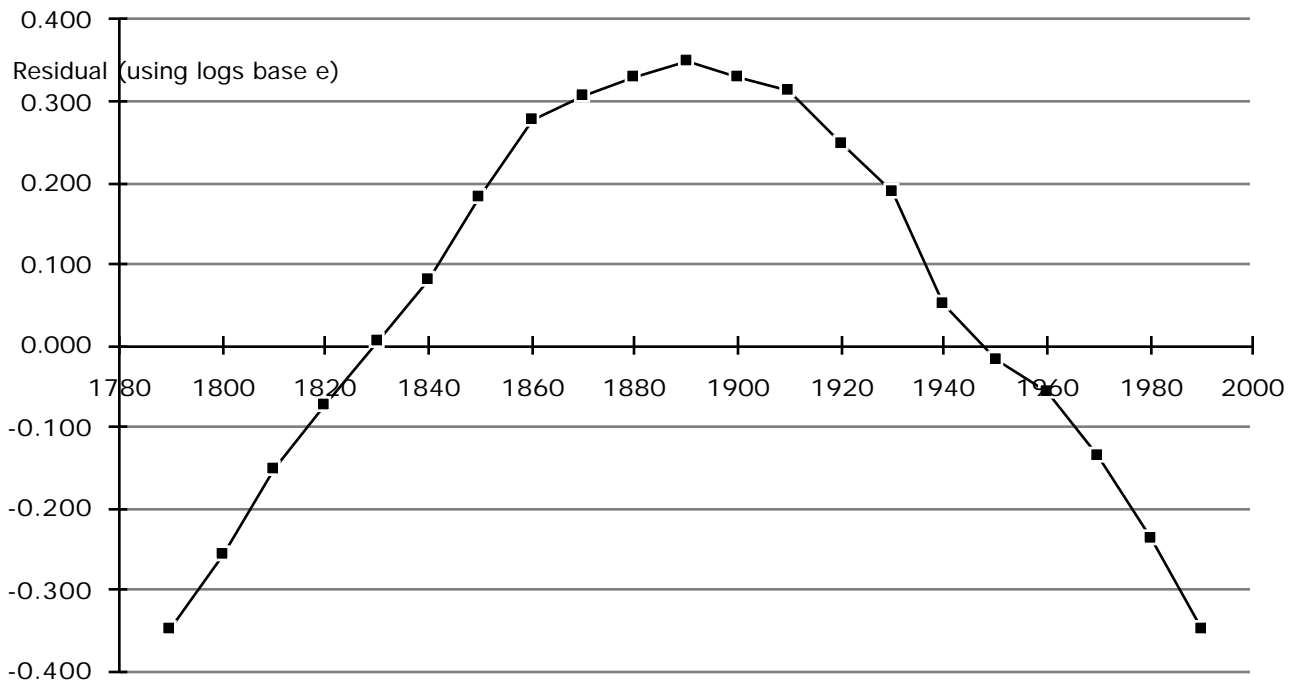
Mean Sqd Resid 0.21243886



There is an obvious problem with this first intercept. So revising the hypothesis by adding 0.35 to the intercept:

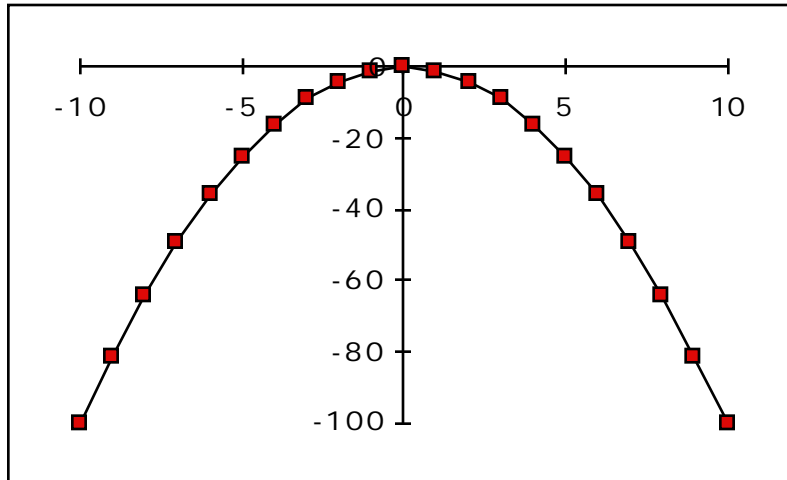
Intercept 15.534  
 slope 0.02073924  
 0

Year	Years after 1790	In(Pop)	Expected	Residual	Sqd Residual	
1790	0	3,929,214	15.184	15.534	-0.350	0.123
1800	10	5,308,483	15.485	15.741	-0.257	0.066
1810	20	7,239,881	15.795	15.949	-0.154	0.024
1820	30	9,638,453	16.081	16.156	-0.075	0.006
1830	40	12,866,020	16.370	16.364	0.007	0.000
1840	50	17,069,453	16.653	16.571	0.082	0.007
1850	60	23,191,876	16.959	16.778	0.181	0.033
1860	70	31,443,321	17.264	16.986	0.278	0.077
1870	80	39,818,449	17.500	17.193	0.307	0.094
1880	90	50,155,783	17.731	17.401	0.330	0.109
1890	100	62,947,714	17.958	17.608	0.350	0.122
1900	110	75,994,575	18.146	17.815	0.331	0.109
1910	120	91,972,266	18.337	18.023	0.314	0.099
1920	130	105,710,620	18.476	18.230	0.246	0.061
1930	140	122,755,046	18.626	18.437	0.188	0.035
1940	150	131,669,275	18.696	18.645	0.051	0.003
1950	160	151,011,580	18.833	18.852	-0.019	0.000
1960	170	179,143,706	19.004	19.060	-0.056	0.003
1970	180	203,302,231	19.130	19.267	-0.137	0.019
1980	190	226,545,805	19.238	19.474	-0.236	0.056
1990	200	248,709,873	19.332	19.682	-0.350	0.123
				Mean Resid	Sqd	0.05557615



Clearly, these residuals show a pattern and, therefore, the hypothesis is wrong. The fit of this line to the logs yields residuals ranging from -.35 to +.35, meaning that the ratio between the true populations and the populations that would be expected (were the hypothesis correct) range as high as the exponential of .35, which is 1.41, errors of 41% at the extremes.

But, look at the graph. It looks “sort of” quadratic. It looks like an almost straight line, rising. It bends. And then it looks almost straight, falling away. My mathematical repertoire tells me that quadratic equations can look like that. For example, here is a graph of the function  $y = -x^2$ , graphed between  $x = -10$  and  $x=10$ .



Is that mathematical pattern the pattern I've seen in these residuals? Let me hypothesize that it is, and then test it.

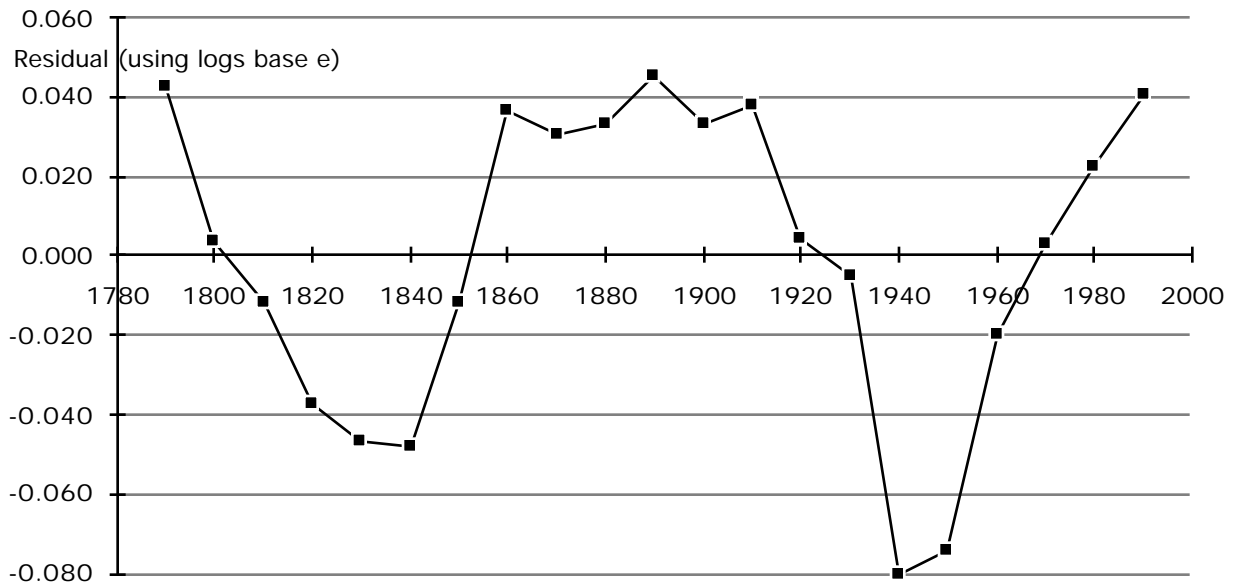
But first, never mind the mathematics, what would it mean if the hypothesis were correct? How would I interpret a quadratic equation and what would it tell me? My reflex is to search my memory for something analogous. And what I come up with is something from simple physics: The equation for the position of a particle moving in a straight line. If the particle begins at  $x_0$  with velocity  $v$  (at time 0), and accelerates with acceleration  $a$ , then the descriptive equation is

$$x(t) = x_0 + vt + \frac{1}{2} at^2$$

That feeds my intuition: If a quadratic equation fits these data, then the coefficient of the quadratic term would be describing acceleration or deceleration in the growth rate. That's good. I know already that the growth rate in 1790 is about one percent larger than the growth rate in 1990. The quadratic equation would express a hypothesis that the growth declined smoothly over the two hundred periods (not suddenly, circa 1890). So, I will be able to interpret it, if I need to.

So, back to my spread sheet, I keep adjusting my estimates of the coefficients, reducing the mean squared residual and I get:

			Intercept	15.1411		
			slope	0.0346704		
			Quadratic	-6.96E-05		
Year	Years after 1790		ln(Pop)	Expected	Residual	Sqd Residual
1790	0	3,929,214	15.184	15.141	0.043	0.002
1800	10	5,308,483	15.485	15.481	0.004	0.000
1810	20	7,239,881	15.795	15.807	-0.012	0.000
1820	30	9,638,453	16.081	16.119	-0.037	0.001
1830	40	12,866,020	16.370	16.417	-0.046	0.002
1840	50	17,069,453	16.653	16.701	-0.048	0.002
1850	60	23,191,876	16.959	16.971	-0.011	0.000
1860	70	31,443,321	17.264	17.227	0.037	0.001
1870	80	39,818,449	17.500	17.469	0.031	0.001
1880	90	50,155,783	17.731	17.698	0.033	0.001
1890	100	62,947,714	17.958	17.912	0.046	0.002
1900	110	75,994,575	18.146	18.113	0.033	0.001
1910	120	91,972,266	18.337	18.299	0.038	0.001
1920	130	105,710,620	18.476	18.472	0.004	0.000
1930	140	122,755,046	18.626	18.631	-0.005	0.000
1940	150	131,669,275	18.696	18.776	-0.080	0.006
1950	160	151,011,580	18.833	18.907	-0.074	0.005
1960	170	179,143,706	19.004	19.024	-0.020	0.000
1970	180	203,302,231	19.130	19.127	0.003	0.000
1980	190	226,545,805	19.238	19.216	0.022	0.001
1990	200	248,709,873	19.332	19.291	0.040	0.002
					Mean Resid	Sqd 0.00144635



That is interesting. My first inspection sees two things: The residuals range from  $-0.08$  to  $+0.04$ , which means that the population estimates are off by a maximum of 4 to 8 percent. Second, the residuals are flat circa 1890 — which is the point at which the earlier analysis hypothesized a break in the pattern.

In more detail, I see that the errors are, many of them, on the order of 4%, slightly larger but not dramatically larger than the errors visible in the earlier analysis. So, this description is “competitive” with the earlier analysis.

Second, I am worried by the appearance of “cycling” in these residuals. Is this the signal I have just warned myself about? If fitting a line leaves residuals that are quadratic, if fitting a quadratic leaves residuals that are quartic — then you probably should stop and think rather than proceeding ever further through an infinite regress. But that is not what happened here. Here, I fit a quadratic and got residuals with one peak (the middle) and two valleys, that's one more than I would have gotten if I were locked into a polynomial regress. That doesn't prove that I'm

not in trouble. But it is re-assuring. These residuals may, in fact, be real — not the product of a misguided analysis.

You should also note that the comparison between these two analyses shows an example of a general rule that appears to be completely counter-intuitive: You look at a graph and see various bumps and curves. The mind seizes on these features, and the scientist within says: “How can I explain that?” The bumps and patterns look real. They are the “that” that needs to be explained. But it isn’t that simple. In most cases the bumps and patterns are bumps and patterns as compared to some base-line expectation. And if you change the base line expectation, this thing in your head which you make appear on the graph, then you change the bumps and patterns. The mind seizes on these alternative feature and the scientist within goes off in another direction.

With these data the first analysis seized on a bend in the curve, circa 1870. On the graph of residuals for the first 100 years, this point for 1870 was approximately 6% below the residual for 1860. It attempted to “explain” that by breaking the data into two batches — corresponding to an idea that the process itself had changed circa 1870. So the analysis said “explain” — wrap a story around — exponential growth at one rate during the first 100 years, a change in the process, and then exponential growth during the remaining years. The raw material calling for this explanation was the bump. The explanation will focus on some relatively sudden shift, circa 1870.

The second analysis compared the whole (not the parts) to a process in which an initial growth rate of about 3% tapered off slowly and smoothly during the ensuing 200 years, down to a growth rate of about 2%. That reference shifts the attention. Now the interesting remaining features are the descents into two valleys and the climbs out of them, two valleys separated by about 100 years. The least interesting part of the graph is the middle: the residuals for 1860 through 1910 are all up about 4%, with nothing remarkable demanding attention to 1870. Compared to a basically constant growth rate, modified by a very slow decline (from 3% to 2% in 200 years), the middle of the curve is flat and uninteresting. The hypothesis handles the middle years very nicely, shifting attention to the two valleys, circa 1840 and circa 1940. The point is that the very phenomenon that we think we have to explain is already, in part, a product of the analysis. How do you reduce the subjectivity that is built into what appear to be facts? By extreme skepticism, by constant testing, by enlarging the research to data that can show

themselves to be consistent with a hypothesis, thus supporting it, or inconsistent with a hypothesis, thus rejecting it. We also have rules, like choosing parsimony: In this case, the first description requires two straight lines separated by a break while the second description requires one quadratic equation, with no break. The second is a more parsimonious description. I also have a rule of skepticism when I see things that are too regular. In this case the residuals are too symmetrical, displaying mirror symmetry around the middle. That suggests that there may be a better single equation, that the one I've used is wrong, and that these residuals are a mathematical result of the difference between the better equation and the one I've used.

So, what have I established and what can I speculate

<p>The population of the United States increased from 3.9 million to 250 million during the 200 years from 1790 to 1990.</p>	<p>✓ Fact</p>																																										
<p>The rate of increase has declined from about 3 percent per annum to about 1 percent per annum.</p>	<p>✓ Fact</p>																																										
<p style="text-align: center;">Change (in logs) compared to previous census</p> <table border="1"> <caption>Estimated data points from the graph</caption> <thead> <tr> <th>Year</th> <th>Change (in logs)</th> </tr> </thead> <tbody> <tr><td>1800</td><td>0.30</td></tr> <tr><td>1810</td><td>0.31</td></tr> <tr><td>1820</td><td>0.29</td></tr> <tr><td>1830</td><td>0.29</td></tr> <tr><td>1840</td><td>0.29</td></tr> <tr><td>1850</td><td>0.30</td></tr> <tr><td>1860</td><td>0.30</td></tr> <tr><td>1870</td><td>0.24</td></tr> <tr><td>1880</td><td>0.23</td></tr> <tr><td>1890</td><td>0.23</td></tr> <tr><td>1900</td><td>0.19</td></tr> <tr><td>1910</td><td>0.19</td></tr> <tr><td>1920</td><td>0.14</td></tr> <tr><td>1930</td><td>0.15</td></tr> <tr><td>1940</td><td>0.07</td></tr> <tr><td>1950</td><td>0.14</td></tr> <tr><td>1960</td><td>0.17</td></tr> <tr><td>1970</td><td>0.13</td></tr> <tr><td>1980</td><td>0.11</td></tr> <tr><td>1990</td><td>0.10</td></tr> </tbody> </table>		Year	Change (in logs)	1800	0.30	1810	0.31	1820	0.29	1830	0.29	1840	0.29	1850	0.30	1860	0.30	1870	0.24	1880	0.23	1890	0.23	1900	0.19	1910	0.19	1920	0.14	1930	0.15	1940	0.07	1950	0.14	1960	0.17	1970	0.13	1980	0.11	1990	0.10
Year	Change (in logs)																																										
1800	0.30																																										
1810	0.31																																										
1820	0.29																																										
1830	0.29																																										
1840	0.29																																										
1850	0.30																																										
1860	0.30																																										
1870	0.24																																										
1880	0.23																																										
1890	0.23																																										
1900	0.19																																										
1910	0.19																																										
1920	0.14																																										
1930	0.15																																										
1940	0.07																																										
1950	0.14																																										
1960	0.17																																										
1970	0.13																																										
1980	0.11																																										
1990	0.10																																										

<p>Relative to these long term trends there have been two short term increases in the rate of increase, one following 1840, the other following 1940.</p>	<p>- Relative statement, true in stated context.</p>
<p>The first increase may be due to massive immigration following the railroad expansion into farm land of the "West" and the post World War II baby boom.</p>	<p>? Speculation. Really unacceptable speculation, were this a final report, because both statements point to other data that could have been presented but have not been. In the first case, immigration data can be checked to see whether or not it changes circa 1840 and whether or not its magnitude is sufficient to account for the bump in the residuals. In the second case, birth rates can be checked to see whether these birth rates can account for the bump and whether changes in the birth rate were sufficient to create the observed change in the rate of increase for the total population (of all ages).</p>