

# U.S. Population: Not the Work, Not the Report, but the Thinking

## I

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. 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
		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.

---

I am thinking: Populations grow exponentially, which means that each year's growth is proportional to the preceding year's population. How do I know that? Truth is, I don't. But that is what all sorts of Malthusian folklore babbles about, so when I look at a population, I think growth rate (percentage) and think of the summary as the average growth rate. It takes people to make people, so at any time the growth of a population "should be" proportional to the size of the population. Do I believe that? No. I'm skeptical. If it were really obvious, if data behaved as data are "supposed to behave", there would be no need to analyze it. That's my thinking about "process" — a rudimentary hypothesis: growth in proportion to size.

That being said, I can make the first rough estimate in my head: The population doubled about six times in two hundred years. Doubling six times in two hundred years implies doubling one time in about 33 years, if the process was constant. And doubling in 33 years implies an annualized rate of increase of about 2% per annum. There is my first description, untested, of the growth process and growth rate for the United States. ("Rule of 70:" Divide 70 by the rate to estimate doubling time. Or, divide 70 by the number of years to get the rate. So,  $70/33 \approx 2.12$ : the population is doubling at about 2% per annum).

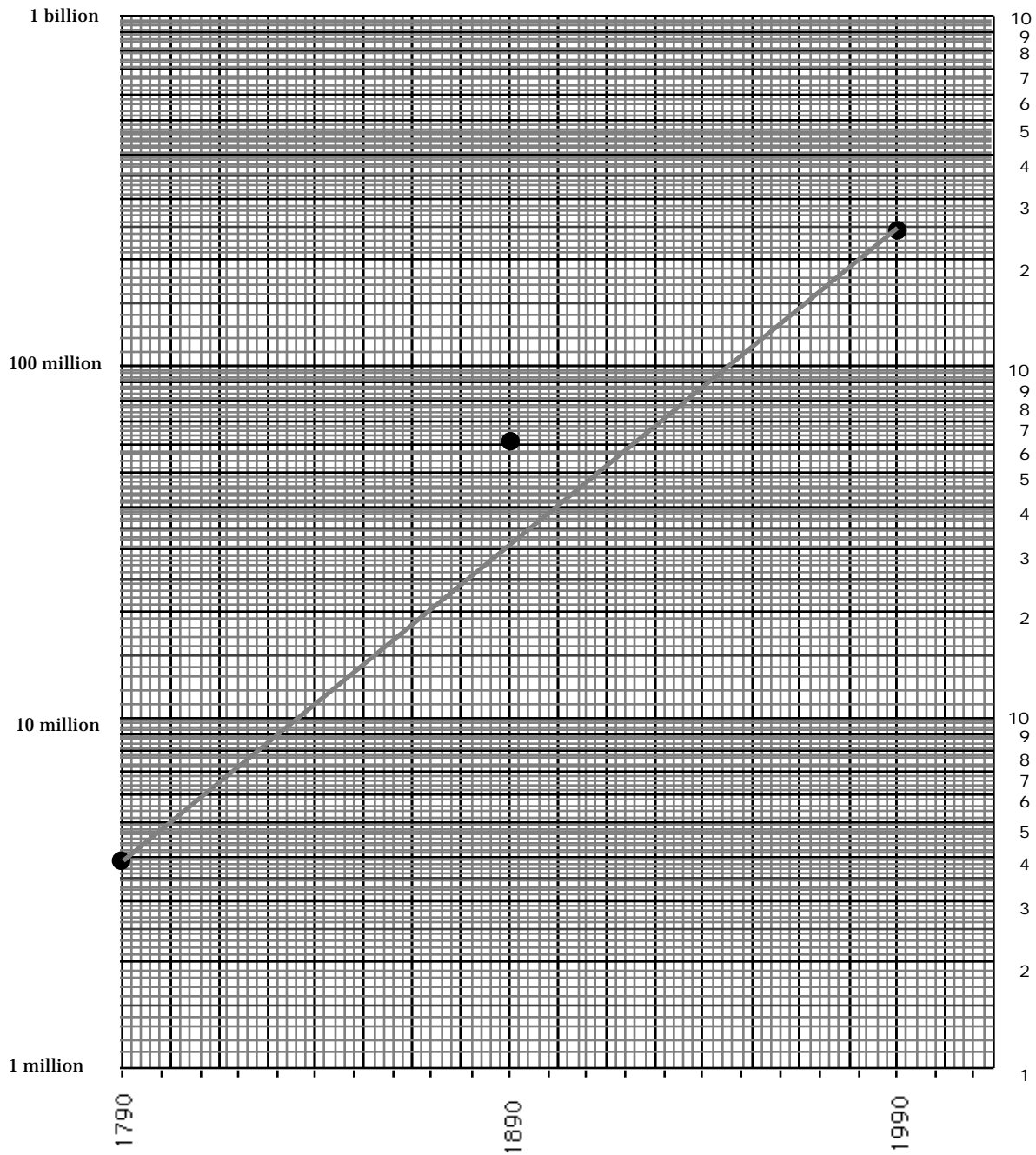
That's my first estimate. It gives me an order of magnitude to think about for all the rest: The annual rate of increase is about 2% per annum. Next, I'll graph it.

What kind of a graph? *Because* I'm thinking "grows by the same percentage each year", I want a kind of graph that is capable of falsifying this hypothesis if the hypothesis is false. I want a kind of graph that will look linear if that hypothesis is correct — but that will look non-linear if the hypothesis is false. If it is false, that will lead me back to re-examine the hypothesis "grows by the same percentage each year". Note: I'm not graphing because I like graphs, nor because that's "the next step". I'm graphing in order to see the actual rate (percentage growth) and I'm graphing so that if it is not behaving that way, then the graph will make it obvious.

So, again, what kind of a graph? Graphing it on ordinary graph paper (graphing the population counts) would be irrelevant — that kind of graph has nothing much to do with what I just said: On ordinary graph paper the growth of the U.S. population is surely going to be non-linear — whether or not my simple hypothesis is correct. So plotting the graph of U.S. population on ordinary graph would not advance my knowledge relative to my hypothesis — an ordinary graph would show that I was not thinking or not thinking clearly. — Waste of time.)

Ah, by contrast: Graphing it on semi-log graph paper, a straight line in *the semi-log graph* would show that my idea was consistent with the data — and failure to find a straight line would show inconsistency — I could learn something from that. So, using semi-log paper, and looking for a line will teach me something relative to my hypothesis.

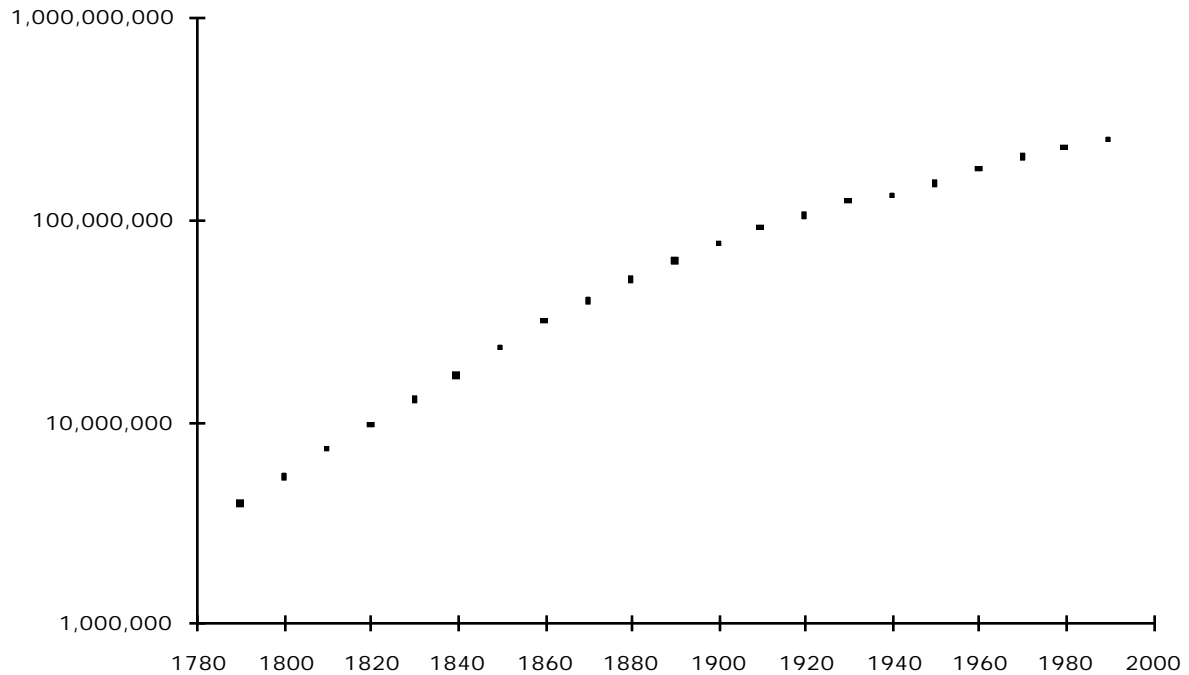
So, semi-log: And truth is, I can learn most of what I want to know by graphing exactly three points: One data point at the left. One data point at the right. And one data point in the middle. The data point in the middle provides the quickest way to test whether or not the data follow a straight line in their course between the first data point and the last. So



Oops: Those three points are not co-linear, not even close. The straight line tells me that my idea, constant proportional growth, would lead to a population of 30 million in 1890. But the data show 60 million in 1890— off by a factor of 2. So the idea is wrong — and I’ve learned something. Thomas Malthus can tell us that population grows exponentially while resources grow linearly, meaning that human populations will outrun the resources that feed us — with disastrous results. And lots of people can think about the great end-of-the-world implications that follow from Malthus’ proposition. *Meanwhile* — we’ve checked the first part of his proposition against the U.S. data And? Its not true: I had an idea, a hypothesis. I framed it in a falsifiable way. And it was false.

One hypothesis — gone. Now, back to thinking. *Idea*: Could I be making too much out of a single point? No, implausible: The hypothetical value, assuming a constant rate of exponential growth missed the true value by a factor of 2, I can’t rescue that idea by invoking “variability”. *Idea*: Real human populations grow and shrink by immigration and emigration, as well as by biological reproduction. I might want to find separate data on these processes. The immigration idea sounds good, although it would require more data — easily obtained. *Idea*: Forsaking Malthus, I remember something about “demographic transition”. That’s what is supposed to separate the “first world” from the “third world”: After industrialization, and a little taste of prosperity (like eating regularly and seeing health of children improve so that they can survive to become adults), people reduce their family size and increase the age at which they begin to bear children. After industrialization there is “supposed to be” a transition. Birth rates are “supposed to” drop — presumably because children get re-classified. Children change from being an asset (as a source of free labor to parents who live off the land) to being a drag on their parents’ income (which is derived from employment away from the household).

That gives me a few ideas to work with, too many perhaps, and for now I’m going to just look at the graph: This time I am just “fishing” to see if it gives me ideas.

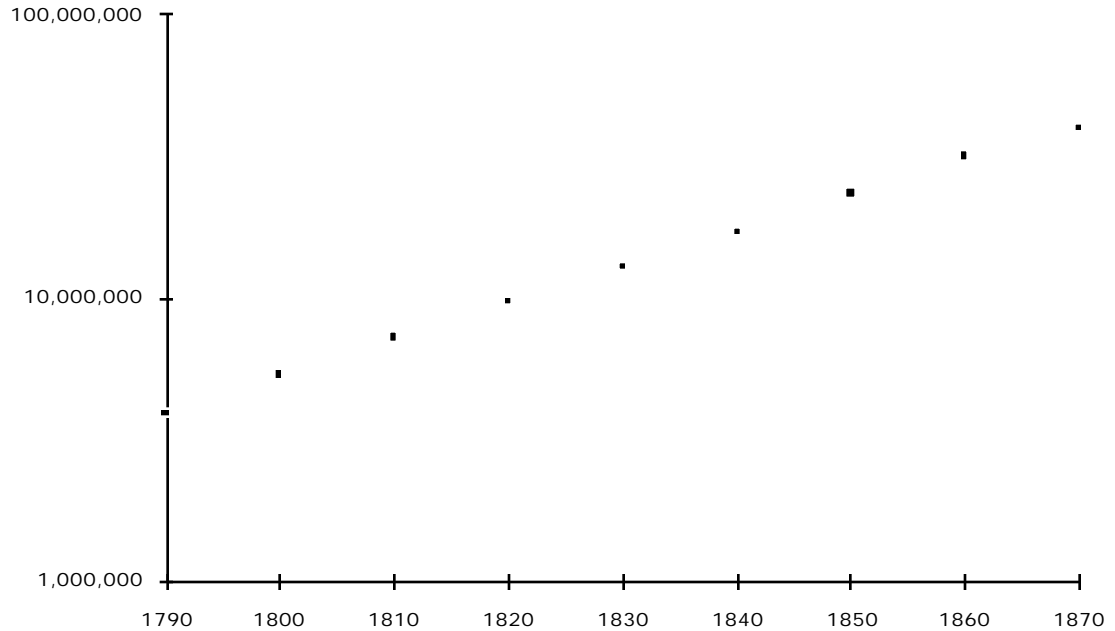


Well, it looks “sort of” linear through the first half. It bends. And then it looks “sort of” linear in the second half at a lower rate (with a dip in 1940). Was there a “demographic transition” between the first half and the second? Could be: U. S. industrialization is really supposed to have “taken off” in the 1870’s to 1890’s — transcontinental railroads, unified markets, telegraph, steel, oil, cartels, ... Let me start easy: Does the first half show a constant annual rate, disrupted in the late 19th century? That will not give me a falsifiable statement about the relatively sophisticated idea of a “demographic transition.” But I needn’t jump that far this quickly: I can put that off while I check the simpler statement “constant annual rate, disrupted in the late 19th century”.

So let me try 1790 through 1870. I want to graph it again. First I’ll do the semi-log graph. Why? to get a close-up” looking for non-linearity. If it is still plausible that the relation is linear, then I’ll look at

the graph of logs to get an estimate of the slope and intercept. And then I'll switch from the logs to the residuals. That's what I really want to see: the residuals, asking: "Do the residuals (for the early years) show a serious departure from linearity (on the semi-log graph)?"

So, I'll get the easy close-up traced on 2 cycle semi-log paper.



Looks good enough. Now, I'll invest the time to compute the logarithms numerically, computing the logarithms that were done automatically, or implicitly, by the semi-log graph paper. Why use the numbers? Because I want to see the residuals and to "see" them I have to compute them. And since I am going to put numbers on the logs, which logs? I'll use natural logs, logarithms base  $e$ , because, with base  $e$ , it is easy to recognize annual rates like 1 and 2 percent — which come out as .01 and .02 in logs base  $e$ .

So, setting-up my computation

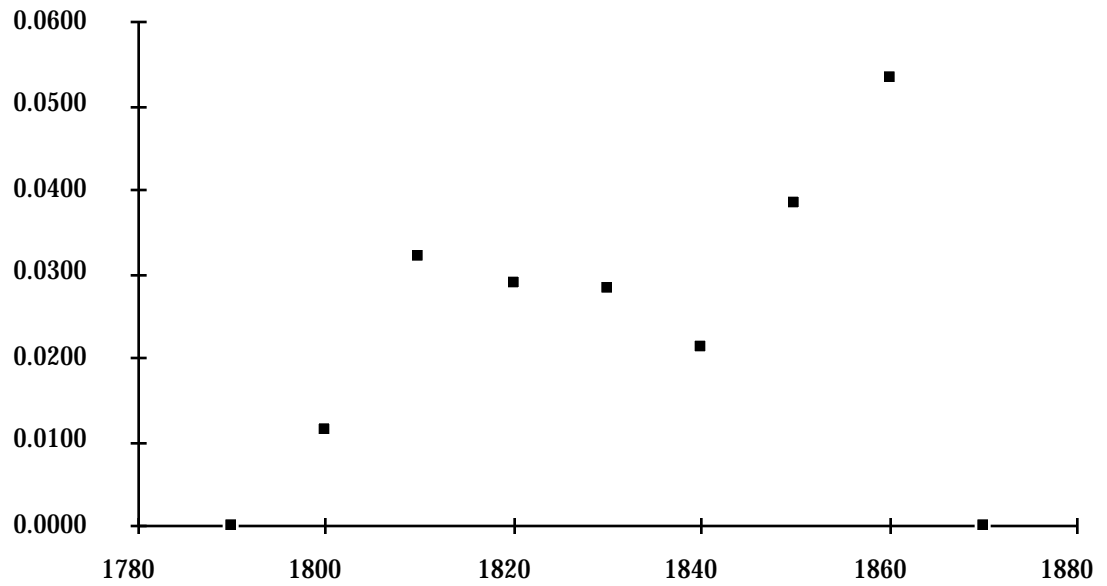
	slope	0.028948637		
	intercept	15.1839		
		Linear Prediction	Residual	Absolute Resid
1790	15.1839	15.1839	0.0000	0.0000
1800	15.4848	15.4734	0.0114	0.0114
1810	15.7951	15.7629	0.0322	0.0322
1820	16.0813	16.0524	0.0289	0.0289
1830	16.3701	16.3418	0.0283	0.0283
1840	16.6528	16.6313	0.0215	0.0215
1850	16.9593	16.9208	0.0385	0.0385
1860	17.2637	17.2103	0.0534	0.0534
1870	17.4998	17.4998	0.0000	0.0000
			Average	Average:

How to estimate slope and intercept? No problem: I just need a rough sketch. So, for a first estimate:

Slope: Last value minus first value, divided by 80.

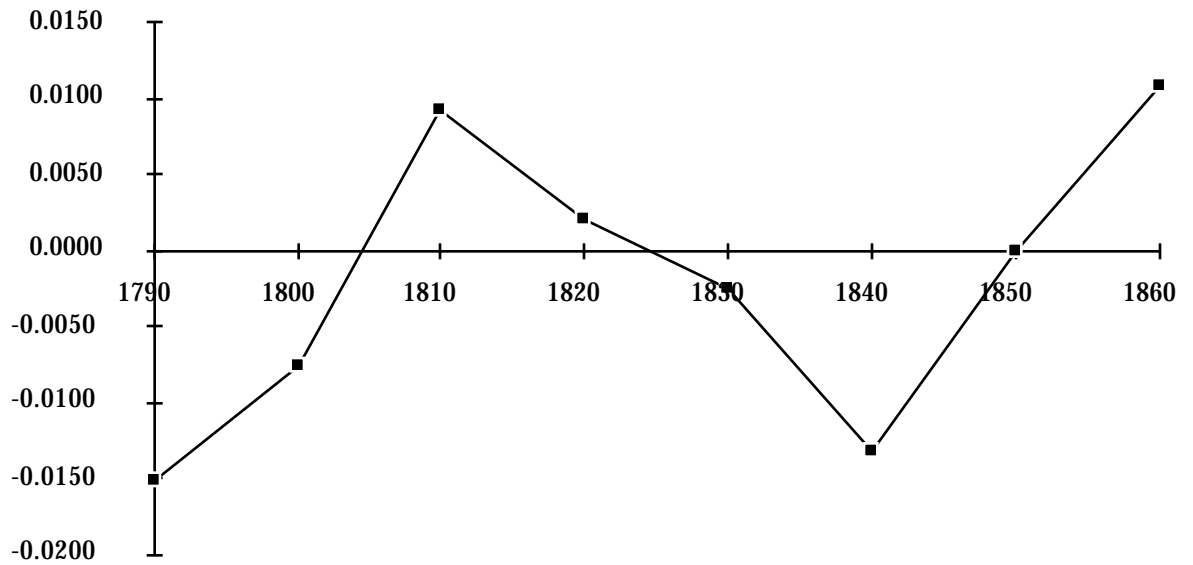
Intercept: I recalibrate "x" to be year since 1790. So "Year 0" corresponds to 1790. Then my estimate for the intercept is First value.

There's the graph. Strange: all the residuals are positive, and they are zero at both ends. That's a warning of curvature, but I'd better look more closely.



It surely isn't simple curvature. Maybe the last point, for 1870, is already out of the range for the early years and entering the range of industrialization. So let me drop the 1870 data point, and commit myself to the work involved in getting a more a serious attempt to get the slope and the intercept. I zero-in on it by looking at both the average residual and the average absolute residual: I fix the intercept so that the average residual is close to zero. Then I fix the slope to make the average *absolute* residual small. Then I fix the intercept so that the average residual is close to zero. Then I fix the slope to make the average *absolute* residual small. Then I fix the intercept so that the average residual is close to zero. ....

Year	Observed (In natural logarithms.)	Linear Prediction	Residual	Absolute Residual	Rank of Absolute Residuals
1790	15.1839	15.1990	-0.0151	0.0151	8
1800	15.4848	15.4924	-0.0076	0.0076	4
1810	15.7951	15.7858	0.0093	0.0093	5
1820	16.0813	16.0792	0.0021	0.0021	2
1830	16.3701	16.3726	-0.0025	0.0025	3
1840	16.6528	16.6660	-0.0132	0.0132	7
1850	16.9593	16.9594	-0.0001	0.0001	1
1860	17.2637	17.2528	0.0109	0.0109	6
1870	17.4998	17.5462	-0.0464	0.0464	9
	slope	0.02934			
	intercept	15.19900	Average Residual:	Average Absolute Residual:	Median Absolute Residual
			-0.002016972	0.007587886	.0093
	exptl of slope	1.02977	Exponentiated:	1.0076	1.0093
	exptl of inter	3,988,796			



Oops, that bothers me: That kind of cycling is exactly what you get when you are thinking about something wrong. (Although it can also be what you get from random numbers. I'm worried that while I have chosen to think about it in logs, my choice of this log form may not be valid. And why logs? Which is to say, why proportions? Do I really believe that populations grow in proportion to the number of people in the population? On second and third thought, prompted by the facts, I'm not so sure about that. Proportional to the number of women is probably closer to the mark, but just barely: The point being that people are not yeast, one indistinguishable from another, all happily reproducing in the presence of nourishment and warmth. Human populations don't grow like that. A large part of the human population is not even "at risk" for reproduction, indeed the most rapidly increasing age cohorts of the population (the older cohorts), are little involved in reproduction. Got to think more carefully, about "proportional growth". Actually, now that I think about it, I'll bet that if we imagine that humans could live forever, our numbers would not increase out exponentially at all — the population growth is more likely to look linear, as the child bearing population size becomes stable in size (and a steadily decreasing part of

the total population. So, “populations grow in proportion to their present size” is sloppy thinking. More accurate to say, we customarily *measure* the growth of populations by reporting the growth in proportion to previous size. Whether or not that growth rate is constant, or whether the growth is proportional to size — those are empirical questions.

For the moment, O.K., whatever that is, whether it is a cycling of some sort, or just an up and down — I have stripped the signal present in these data so that what’s left, the residuals, is small — perhaps too small to support my complicated theories that might have been built on them: How large are the residuals? The magnitudes of the residuals are, at worst, about 1.5 percent compared for these 10th year observations. With an annual growth rate of approximately 2.9% per year — that means my 10th year observations are off by a fraction of a single year’s growth. That seems small. So, summing it up: 1790 to 1860, the average annual growth rate is about 2.98 percent. Applied to the seventy year period of these data, this growth rate predicts the population with a median error of less than 1% (using the median absolute residual, .0093, exponentiating it to 1.0093, converting it to a percent at 0.9%, and reporting it as approximately 1%). There is a suggestion of cycling, relative rapid growth, 1790 to 1820, relative slow, 1820 to 1870, then up. (Editing myself again: I’d Better get some number on those ups and downs. What numbers? People think in terms of annualized rates, so I want the annualized rates corresponding to these ten year periods. So, I compute the ten-year ratios, later to earlier, e.g., the ratio, 5,308,483 to 3929,214 for 1790 to 1800. That’s the multiplier for those 10 years. Then I estimate the annual ratio from the 10 year ratio by computing the 1/10 th power of the ten year ratio. And that’s the multiplier for the average single year, among the 10. That’s 1.0305.

1790	3,929,214	
1800	5,308,483	1.0305
1810	7,239,881	1.0315
1820	9,638,453	1.0290
1830	12,866,020	1.0293
1840	17,069,453	1.0287
1850	23,191,876	1.0311
1860	31,443,321	1.0309

So, 1790 to 1820, about 3.0%. 1820 to 1840, about 2.9% Those are the top and the bottom of one of the “cycles” I was seeing on the graph of the residuals That’s small stuff (small difference) Maybe it is a pattern, it “looks” that way, but the difference between the rapid, 1790 to 1820, and the slow, 1820 to 1840, is too small for me to worry about, a difference of maybe one tenth of one percent, about 10,000 people on the 1820 population of 10 million.

That’s my thinking. Without the thinking there is no reason for me to choose one graph or one set of numbers in preference to some other graphs and numbers. Without the thinking there is no reason for logs or not logs, for residuals, or something else.

---

Exercise:

Complete the Analysis:

Analyze the U.S. Population data for 1870-1890.

Write it up for 1790 - forward.