Modeling population growth in the United States

Luke Paulsen
OpenIntro
openintro.org

1 Introduction

Population growth can define a community. Communities that grow rapidly may see increased investment, while contracting communities may find local assets, such as homes and businesses, falling in value relative to those of nearby communities. In this investigation, we wish to determine which demographic factors relate most closely to population growth in U.S. counties from 2000 to 2010.

2 Data Exploration

Data for each county is available from the US Census website, including age, gender, race, and education, along with other relevant demographics such as homeownership, employment, and income. Five counties for these data are summarized in Table 1, and the data were originally collected from the US Census website.[1] This investigation will only consider a subset of variables and be limited to counties where those variables are complete. The resulting data set represents 3,083 counties on 23 different variables. A complete list of the variables under consideration along with variable descriptions is available at

www.openintro.org/stat/data/cc.php

	growth	pop2000	age_under_5	age_under_18	female	black	hs_grad	bachelors
1	24.96	43671	6.6	26.8	51.3	17.7	85.3	21.7
2	29.80	140415	6.1	23.0	51.1	9.4	87.6	26.8
3	-5.44	29038	6.2	21.9	46.9	46.9	71.9	13.5
4	10.03	20826	6.0	22.7	46.3	22.0	74.5	10.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3143	8.49	6644	5.7	21.8	47.4	0.3	91.1	17.9

Table 1: Five rows from the `countyComplete` data set with 8 of the 23 variables.

*This document is released under a Creative Commons Attribution-ShareAlike 3.0 license.

[1]These data were collected from the US Census website. The data are available in the openintro R package and also as a tab-delimited text file at openintro.org/stat.

A variable called `growth` that represents the population growth rate for each county from 2000 to 2010 is included in Table 1, and this variable represents the response variable for the analysis. This variable is summarized in Table 2 and Figure 3. Growth rates for the ten-year period average 5.4%, with the middle half ranging from -2.2% to 10.4%.

Mean	Median	St. Dev.	IQR	Min	Max
5.42%	3.29%	13.18%	12.70%	-46.6%	110.40%

Table 2: Statistical summaries of population growth in US counties from 2000 to 2010.

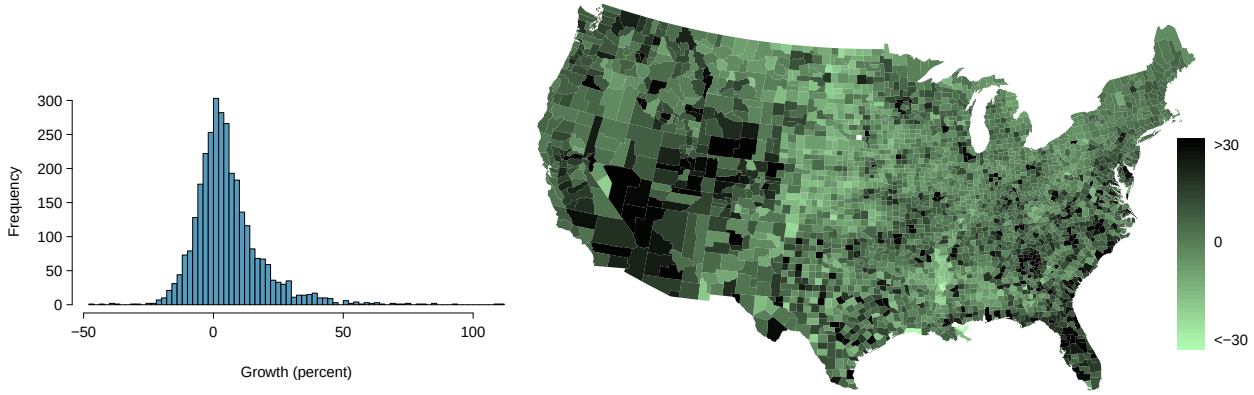

Figure 3: Population growth across the United States from 2000 to 2010.

Several other variables in the data set are worth exploring. The `pop2000` variable measures the population of the county during the 2000 census and is shown in Figure 4. The variable is very skewed, so we will use the natural logarithm of population in the model. Taking the natural logarithm of population allows us to measure population differences in terms of multiplication rather than addition. For example, a difference of 1,000 people would be important in a county with population 10,000 but less so in a county with population 1,000,000. Using the natural logarithm for population means differences are compared geometrically, e.g. comparing counties with populations of 1,000 and 10,000 will be analogous to a comparing two counties with 10,000 and 100,000 people in the model.

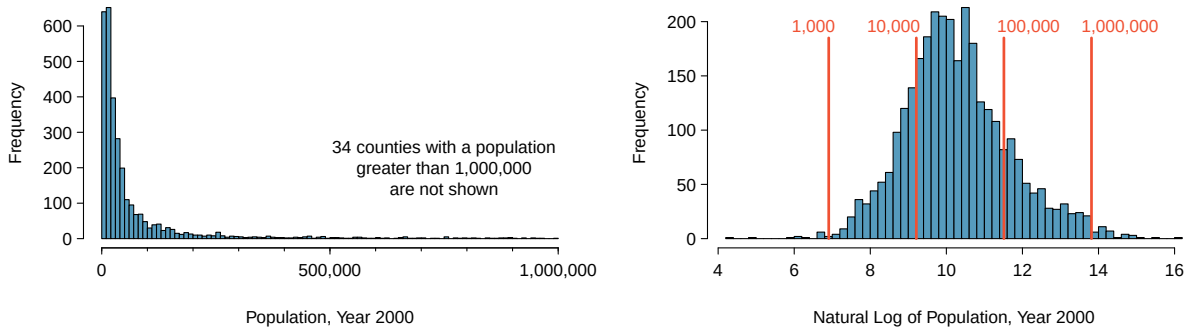

Figure 4: Distribution of populations. Left: original populations. Right: log-transformed populations.

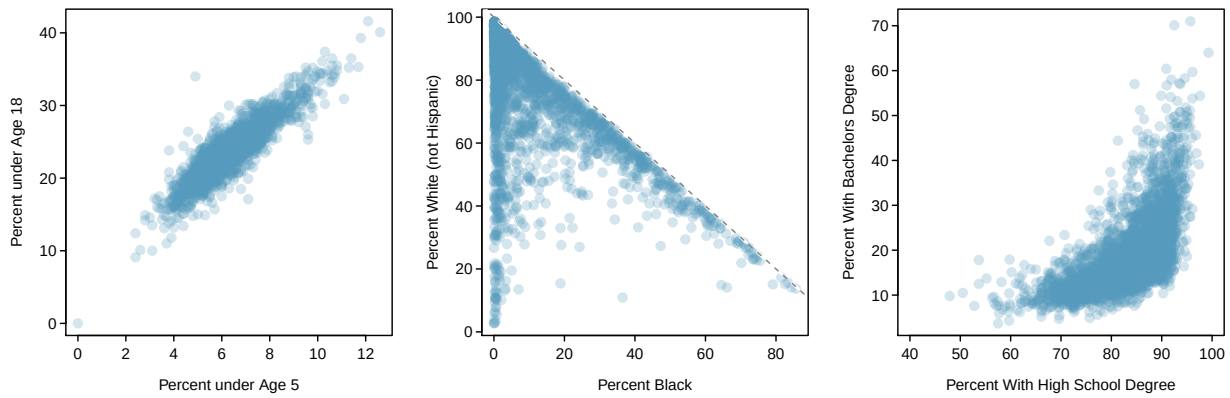

Figure 5: Three figures that highlight the collinearity of several predictors.

There are also several groups of variables that divide the population with respect to a particular statistic: age, race, or education level. We expect these variables to be related to one another, and this relationship must be considered when interpreting the results. Figure 5 highlights the relationships among some of these variables.

The first plot in Figure 5 suggests that the variables `age_under_5` and `age_under_18` are strongly correlated. The diagonal line in the second plot represents the fact that the percentages of each racial group in the population cannot sum to more than 100%. The percentage of the population that self-identifies as some race other than black or non-Hispanic white is represented by the distance of a point from the downward-trending diagonal. The relationship in the third plot is somewhat weaker, but it shows that the percentage of the population with a bachelor's degree is always smaller than the percentage that completed high school, as would be expected.

3 Analysis

Two variables can be linearly related. For example, the left panel of Figure 5 shows a positive trend relating `age_under_5` and `age_under_18`. This trend looks linear, and it can be modeled, even if imperfectly, by using a straight line. Such a line would have error for individual observations, but it would capture the overall structure of the relationship.

3.1 Modeling population growth

When working with many variables, the principles of the linear model can be generalized, where here we simultaneously fit many variables against a response rather than one variable at a time. We begin by writing a formula that models the growth rate as a linear combination of all the other variables that we are considering:

$$\begin{aligned} \widehat{growth} = \beta_0 \ &+ \ \beta_1 \times \text{log(pop2000)} \ + \ \beta_2 \times \text{female} \\ &\vdots \\ &+ \ \beta_{21} \times \text{poverty} \ + \ \beta_{22} \times \text{sales\_per\_capita} \end{aligned}$$

Statistical software may be used to identify the best fitting model, where point estimates of β_0, β_1, ..., β_{22} would be estimated in the model.

To improve the model, we perform model selection, eliminating variables using backwards selection, until all remaining variables are found to be statistically significant. The model following backwards selection is summarized by Table 6.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3168	6.3879	0.83	0.4053
log(pop2000)	1.6556	0.2007	8.25	0.0000
age_under_5	3.3118	0.4179	7.92	0.0000
age_under_18	-0.6160	0.1605	-3.84	0.0001
age_over_65	-0.2666	0.0816	-3.26	0.0011
female	-0.3026	0.1080	-2.80	0.0051
hispanic	0.1320	0.0327	4.03	0.0001
white_not_hispanic	0.0803	0.0145	5.54	0.0000
no_move_in_one_plus_year	-0.5614	0.0520	-10.79	0.0000
foreign_born	0.2509	0.0632	3.97	0.0001
foreign_spoken_at_home	-0.1744	0.0471	-3.70	0.0002
bachelors	0.5281	0.0431	12.24	0.0000
mean_work_travel	0.7156	0.0434	16.48	0.0000
housing_multi_unit	-0.4930	0.0347	-14.22	0.0000
median_val_owner_occupied	0.0000	0.0000	4.20	0.0000
persons_per_household	9.9882	1.3261	7.53	0.0000
per_capita_income	-0.0003	0.0001	-3.73	0.0002
poverty	-0.3266	0.0506	-6.46	0.0000
sales_per_capita	0.0002	0.0000	5.74	0.0000

Table 6: Model summary for the regression model predicting population growth after model selection. See page 2 for a link that provides variable descriptions.

The variables `black`, `hs_grad`, and `density` were eliminated during model selection. However, variables that we would expect to be closely correlated with these variables – `hispanic` and `white_not_hispanic` with `black`, and `bachelors` with `hs_grad` – still appear in the model. As we saw in the Data Exploration section, some variables are highly correlated, i.e. they are collinear. When predictors are collinear, having one in a multiple regression model may be about as good as having both, and this may explain why `black` and `hs_grad` were eliminated during model selection.

In the age variables there is a surprise of a different type. The variables `age_under_5` and `age_under_18` are highly collinear, but both are still included in the model, and the model suggests they have opposing effects on population growth. It may be tempting to make a standard interpretation of the coefficients, however, that could be misleading. These two variables are collinear (see Figure 5), and this complicates interpretation. For example, dropping `age_under_5` results in the coefficient of `age_under_18` changing from -0.62 to 0.40. The practical interpretation of these variables has been complicated by other variables in the model.

3.2 Diagnostics

In order to assess the multiple regression model, we check conditions on the model's residuals. The general requirements are that the residuals are roughly normal, have approximately the same variance, and are independent. We leave it to the reader to check whether any nonlinear relationships exist between the predictors and `growth` variable.

Figure 7 is a normal probability plot of the model's residuals. There is clear curvature, and the tails at the corner of the graph indicate that some of the observations have unusually distant residuals from zero. While this would be a substantial concern for a model with only a small number of data points, over 3,000 counties are being used here, so the influence of these outlying residuals should be very limited.

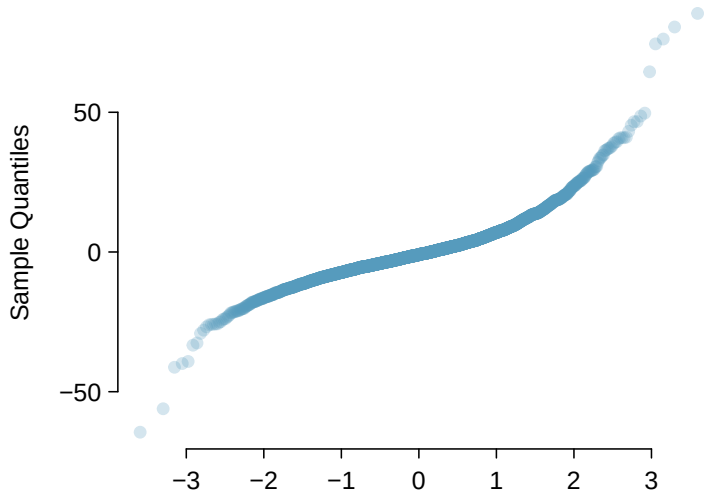

Figure 7: Normal probability plot for the residuals following model selection. There is clear curvature, but the outliers are probably reasonable for the size of this data set.

Figure 8 shows that the residuals plotted against their fitted values. The variance is approximately consistent, with perhaps a small increase in variability with larger predicted values. One county, Kalawao County in Hawaii, had a predicted value far from the cloud at (-59.4%, 20.6%). This small and isolated county was previously a quarantine for leprosy patients; no new residents are allowed to move to this county. Due to the unusual nature of this county, this observation should be excluded in future analyses.

Figure 9 is useful for checking spacial independence of the residuals. In a model that fully explained the observations, we would expect the residual values to be randomly distributed geographically; instead there are definite geographic patterns and clusters of similar residuals. For example, the model fails to account for variables such as climate, which may help explain why adjacent counties tend to have similar residual values. This figure indicates there are additional features remaining within the data that were not captured by the multiple regression model presented in here, violating the independence condition for the residuals.

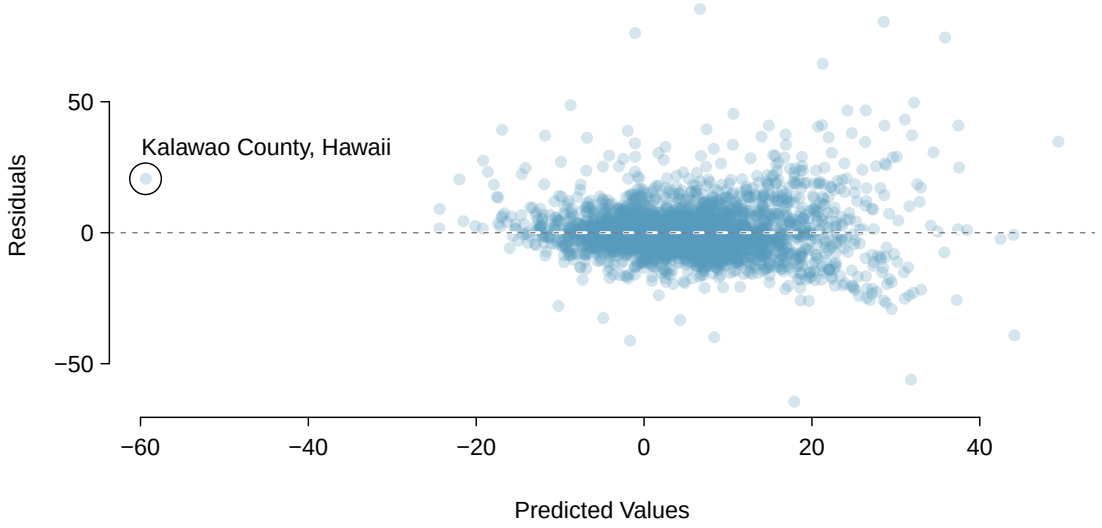

Figure 8: Residuals versus fitted values from the regression model.

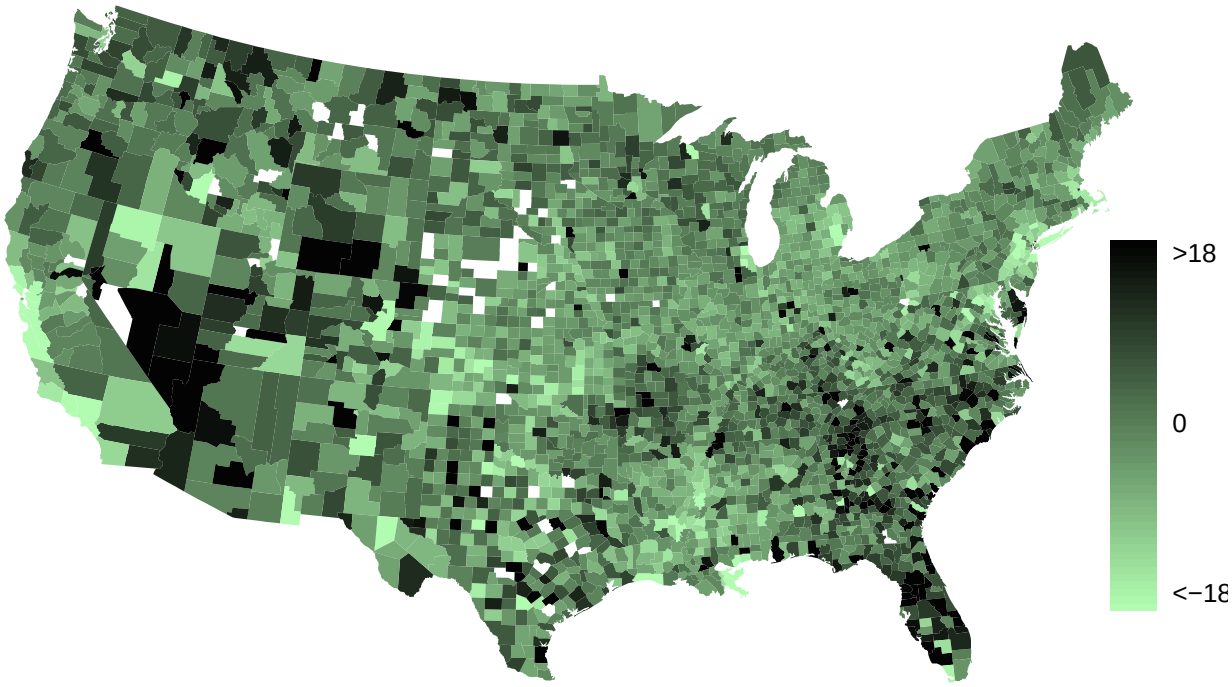

Figure 9: Residuals plotted by their location. Empty spaces represent counties that had missing data and were not included in the analysis.

3.3 Practical interpretation of model coefficients

We will proceed in estimating the impact of many variables on population growth, but we want to highlight that these findings may be somewhat unreliable due to the violation of the independence condition for the residuals. Each variables's coefficient was multiplied by the variable's IQR to get a scaled impact for the variable, shown in Table 10. The proper way to interpret each value is, "The growth rate for a county at the 75th versus the 25th percentile in this variable, other things being equal, would be estimated as ____ higher over ten years."

log(pop2000) 2.8%	age_under_5 4.3%	age_under_18 -2.3%
age_over_65 -1.3%	female -0.4%	hispanic 0.9%
white_not_hispanic 2.2%	no_move_in_one_plus_year -3.2%	foreign_born 1.0%
foreign_spoken_at_home -1.3%	bachelors 5.0%	mean_work_travel 5.1%
housing_multi_unit -4.8%	median_val_owner_occupied 1.1%	persons_per_household 2.6%
per_capita_income -1.9%	poverty -2.6%	sales_per_capita 1.4%

Table 10: The values in this table represent the estimated difference in growth rate for a county at the 75th versus the 25th percentile in each variable, other things being equal.

4 Conclusion

In this investigation, we attempted to model a U.S. county's population growth based on readily available demographic data, a potentially useful tool for economic and other applications. We found strong statistical evidence that many of the demographic variables measured by the 2010 U.S. Census (including age, racial, and demographic distribution, economic conditions, and household makeup) were important in modeling a county's population growth between 2000 and 2010. Taken together in a multiple regression model, the measured variables appear to explain nearly half of the variation in growth rate among counties.

Of the variables measured, the percentage of the population with a bachelor's degree may be especially important in terms of population growth. We suspect the modestly large estimated coefficient of mean commute time is not a driver of population growth but a result of a migration to suburbs, which often require larger commute times. It is also important to consider that many of the variables examined are related to one another, which complicates the interpretability of many model coefficients. This makes it especially difficult to conjecture causal conclusions from the current model.

Further analysis of how the model's variables are related to one another, possibly including transformations of some variables in the model, may be helpful in eliminating this source of error and in providing more definite results. In addition, this model includes no information on geographic

location and does not distinguish between urban, suburban, and rural areas. These types of information appear to be important in determining county growth rates and should likely be included as variables in future investigations.