Evaluating Population Forecast Accuracy:
A Regression Approach Using County Data

Jeff Tayman, UC San Diego
Stanley K. Smith, University of Florida
Stefan Rayer, University of Florida

ABSTRACT

Many studies have evaluated the impact of differences in population size and growth rate on population forecast accuracy. Virtually all these studies have been based on aggregate data; that is, they focused on average errors for places with particular size or growth rate characteristics. In this study, we take a different approach by investigating forecast accuracy using regression models based on data for individual places. Using decennial census data from 1900 to 2000 for 2,482 counties in the United States, we construct a large number of county population forecasts and calculate forecast errors for 10- and 20-year horizons. Then, we develop and evaluate several alternative functional forms of regression models relating population size and growth rate to forecast accuracy; investigate the impact of adding several other explanatory variables; and estimate the relative contributions of each variable to the discriminatory power of the models. Our results confirm several findings reported in previous studies but uncover several new findings as well. We believe regression models based on data for individual places provide powerful but under-utilized tools for investigating the determinants of population forecast accuracy.

INTRODUCTION

Population projections at the state and local levels are used for a wide variety of planning, budgeting, and analytical purposes. Although they are sometimes used simply to trace out the implications of a particular set of hypothetical assumptions, they are used most frequently as forecasts of the future population. The importance of the purposes for which these forecasts are used— for example, opening a new business, closing a public school, enlarging a power plant, or revising local bus routes—makes it essential to evaluate their precision and bias.

Many studies have investigated the impact of population size and growth rate on forecast accuracy by analyzing forecast errors within broad size and growth rate categories. Measuring population size in the launch year and growth rate over the base period, these studies have generally found precision to improve with increases in population size and decline with increases in the absolute value of the growth rate (e.g., Keyfitz 1981; Rayer 2008; Smith and Sincich 1992; Stoto 1983; White 1954). They have found bias to have little or no relationship with population size but to be positively related to the growth rate (e.g., Isserman 1977; Rayer 2008; Smith 1987; Tayman 1996). These results have been found so frequently that we believe they can be accepted as general characteristics of population forecast errors.

All of the studies cited above were based on aggregate data; that is, they focused on average errors for places with particular size or growth rate characteristics. It is clear, however, that errors often vary substantially within given size or growth rate categories. For example, although small places have larger errors than large places on average, some small places have very small errors and some large places have very large errors. What is the impact of population size and growth rate on forecast accuracy when analyzed using data for individual places? What

variables other than size and growth rate also affect accuracy? How can these effects best be evaluated? These questions have seldom been addressed in the literature.

In this study, we investigate population forecast accuracy from a disaggregate perspective. Using a data set covering 2,482 counties in the United States for each census year from 1900 to 2000, we construct county population forecasts and calculate forecast errors for 10- and 20-year horizons.[1] Then, we develop and evaluate several alternative regression models in which population size and growth rate are used as explanatory variables and forecast error as the dependent variable. We extend the analysis to include three additional explanatory variables: prior forecast error, geographic area (defined here as census divisions), and launch year. Finally, we estimate the relative contribution of each explanatory variable to the discriminatory power of the models. The specific questions we address are:

1) Which functional form of a single-variable regression model best describes the relationships between forecast accuracy and population size and growth rate?

2) Can multivariate models improve on the performance of single-variable models?

3) How does the inclusion of prior error, census division, and launch year affect the regression results for population size and growth rate?

4) What are the relative contributions of each explanatory variable to population forecast accuracy?

We are aware of only two studies using regression analysis to investigate population forecast accuracy, one based on aggregate data and one based on data for individual places. Tayman, Schafer, and Carter (1998) used aggregate data to analyze differences in average forecast error by population size and growth rate for a number of small geographic areas in San Diego County. Lenze (2000) used data for individual places but did not evaluate alternative

functional forms of the regression model, did not account for the direction of forecast error, and analyzed only one set of 5-year forecasts for 67 counties in Florida. To our knowledge, this is the first study to evaluate alternative forms of the regression model; to investigate the effects of population size, growth rate, and several other variables on forecast accuracy; to account for both the size and direction of error; and to cover a large number of places, multiple time periods, and several forecast horizons.

Regression analysis illuminates patterns that cannot otherwise be observed and provides a means for testing hypotheses regarding the determinants of population forecast accuracy. We believe the present study provides a fresh perspective on forecast accuracy and deepens our understanding of the factors making some forecasts more accurate than others.

DATA

We conducted our analyses using a data set covering all counties in the United States that did not experience significant boundary changes between 1900 and 2000 (Rayer 2008). This data set included 2,482 counties, 79 percent of the national total. For each county, we collected information on population size in the launch year (the year of the most recent data used to make a forecast), growth rate over the base period (the 10 years immediately prior to the launch year), and forecast errors for 10- and 20-year horizons. The launch years included all decennial census years from 1910 to 1990 for 10-year horizons and from 1910 to 1980 for 20-year horizons.[2]

Forecasts for each launch year were derived from five extrapolation techniques: linear, exponential, share of growth, shift share, and constant share (Rayer 2008). The forecasts analyzed in this study were calculated as an average of the forecasts from these five techniques,

after excluding the highest and lowest. Forecasts refer solely to total population; no forecasts of age, sex, race, or other demographic characteristics were made.

Simple techniques such as these are frequently used for small-area forecasts and have often been found to produce forecasts of total population that are at least as accurate as those produced using more complex techniques (e.g., Chi 2009; Long 1995; Murdock et al. 1984; Rayer 2008; Smith and Sincich 1992; Stoto 1983). An important benefit of these techniques is that they are based on readily available data and can be applied retrospectively to a large data set. Given the similarity of errors when different techniques are applied to the same data set, we do not believe the results reported here are affected by the choice of forecasting technique.

Forecast error was calculated as the percent difference between the population forecasted for a particular year and the population for that year counted in the decennial census. Errors were measured in two ways, one ignoring the direction of error (called "absolute" percent error) and the other accounting for the direction of error (called "algebraic" percent error). The first is a measure of precision and the second is a measure of bias.

Table 1 summarizes population size and growth rate characteristics for counties in the data set. Although mean population size more than tripled between 1900 and 2000, median size increased by only 53%. The 90th percentile population size grew by 278%, but the 10th percentile size grew by only 46%. Mean growth rates were higher than median growth rates in every decade and varied more over time. In most decades, 40-50% of counties lost population. For more detailed information on the data set and forecasting techniques, see Rayer (2008).

(Table 1 about here)

Table 2 provides a summary of the relationships between average forecast errors and the five explanatory variables examined in this study, using discrete categories for each variable and

forecasts covering 10-year horizons.[3] As shown in the top two panels, mean absolute percent errors (MAPEs) had a negative relationship with population size and a u-shaped relationship with the growth rate. Mean algebraic percent errors (MALPEs) had a weak positive relationship with population size and a considerably stronger positive relationship with the growth rate. Similar results were found for 20-year horizons (not shown here). These relationships are consistent with findings reported in many previous studies.

(Table 2 about here)

The next two panels show the relationship between prior error and forecast accuracy. Prior absolute percent errors displayed a strong positive relationship with subsequent MAPEs but no clear relationship with subsequent MALPEs. Prior algebraic percent errors displayed a strong u-shaped relationship with subsequent MAPEs and a strong negative relationship with subsequent MALPEs. That is, prior absolute errors were related to the precision but not the bias of subsequent forecasts, whereas prior algebraic errors were related to both precision and bias. Again, results for 20-year horizons were similar to those shown here for 10-year horizons.

The negative relationship between prior algebraic errors and MALPEs may seem puzzling at first, but it can be understood in light of the finding that extreme growth rates tend to regress toward the mean over time (e.g., Smith 1987). For example, suppose that a county grew unusually rapidly between 1950 and 1960. As a result of this growth, the forecast for 1960 made in 1950 was too low (i.e., the error was negative). Because of the rapid growth between 1950 and 1960, growth was forecasted to be rapid between 1960 and 1970. With regression to the mean, however, growth between 1960 and 1970 was less than forecasted, leading to a positive error in 1970. The negative error for 1960 was thus associated with a positive error for 1970.

Given these patterns, it is not surprising that MALPEs tend to be positive for counties with negative prior errors and negative for counties with positive prior errors.

The final two panels show errors for census divisions and individual launch years. MAPEs were lowest in the Northeast and Midwest and highest in the South and West, but MALPEs showed no clear relationship with the location of census divisions. With respect to launch year, MAPEs followed no clear pattern over time, fluctuating within a narrow range of 9-14%. MALPEs also followed no clear pattern over time, but fluctuated over a considerably wider range (-9% to 9%). Larger year-to-year variations in MALPEs than in MAPEs have been noted previously (e.g., Smith and Sincich 1988).

REGRESSION MODELS AND ANALYSES

Table 2 illustrates the approach followed in most studies of population forecast accuracy; namely, using aggregate data to compare average errors for places with different values of a given characteristic. In this study, we go beyond this approach by constructing regression models based on data for individual counties. We use two data sets, one containing all 10-year forecasts with launch years from 1910 to 1990 and one containing all 20-year forecasts with launch years from 1910 to 1980. Our dependent variables are absolute and algebraic percent errors; these are measures of precision and bias, respectively. We use several explanatory variables and functional forms to construct a number of different regression models.

We started with population size and growth rate as explanatory variables; these are the variables most frequently used in evaluations of population forecast accuracy. We calculated growth rates as absolute values in regressions related to precision and as algebraic values in regressions related to bias; we refer to these variables as GR-Abs and GR-Alg, respectively. We

constructed two simple single-variable models, one with population size as the explanatory

variable and the other with the growth rate. Then, we evaluated alternative functional forms of

each single-variable model and chose the optimal form; we refer to these as complex single-

variable models. Then, we combined size and growth rate in a multivariate model. Finally, we

added prior error, census division, and launch year as explanatory variables in the multivariate

model.

Based on the findings of previous studies and the data shown in Table 2, we hypothesize

that increases in population size will improve precision but have little impact on bias; that

increases in the absolute value of the growth rate will reduce precision; and that increases in the

algebraic value of the growth rate will reduce downward bias in counties losing population and

raise upward bias in counties gaining population (i.e., they will have a positive effect on

algebraic percent errors).

Population size and growth rate are not the only factors affecting forecast accuracy, of

course. Other factors that may be important are economic conditions (e.g., job openings, wage

rates, cost of living), social conditions (e.g., educational opportunities, racial discrimination),

demographic conditions (age structure, ethnic composition), and environmental conditions (e.g.,

pollution levels, water supplies). Although our data set does not contain information pertaining

directly to these factors, we have selected three explanatory variables that may reflect their net

impact: prior error, census division, and launch year.

Prior forecast errors reflect the impact of factors other than population size and growth

rate that make it particularly easy or difficult to forecast accurately. These errors may provide

useful predictors of future errors. For example, the 10-year forecast error for launch year 1950

may provide a useful predictor of the 10-year forecast error for launch year 1960. We calculated

prior errors using the same number of years as were included in the forecast horizon (e.g., for forecasts with a 20-year horizon, we used the error for the 20-year forecast ending in the launch year). As we did with growth rates, we used absolute values of prior errors in regressions related to precision (Prior-Abs) and algebraic values in regressions related to bias (Prior-Alg). Based on the data shown in Table 2, we hypothesize that prior absolute errors will have a positive effect on subsequent absolute errors and prior algebraic errors will have a negative effect on subsequent algebraic errors.

We used census division as a proxy for geographic differences in economic, social, demographic, and environmental conditions and launch year as a proxy for changes in those conditions over time. Census divisions and launch years were measured using a series of dummy variables. For census division, the reference group was the South Atlantic division, a division with size, growth rate, and forecast error characteristics similar to those for the entire U.S. For launch year, the reference group was 1960, near the middle of the century and a year with a moderate forecast error. We do not have any *a priori* expectations regarding the effects of these two variables on precision or bias.

Prior error, census division, and launch year are illustrative of the types of explanatory variables that could be used in regression analyses of forecast accuracy. They are particularly appropriate for the present study because they are available for all counties and decennial census years. Other variables could also be used, of course.

Simple Single-Variable Models

Regression coefficients and adjusted R^2 values for the simple single-variable models are shown in Table 3. The top panel shows the results for absolute percent errors. In every instance, the

explanatory variable had the expected sign and was statistically significant: Increases in population size reduced errors and increases in the absolute value of the growth rate raised errors. However, as shown by the small adjusted R^2 values, neither variable explained much of the variation in forecast errors. Increasing the length of the forecast horizon had little impact on either the regression coefficients or the adjusted R^2 values.

(Table 3 about here)

The bottom panel of Table 3 shows the results for algebraic percent errors. Population size had a significant positive effect on algebraic errors for both forecast horizons, contradicting our hypothesis. However, the coefficients were much smaller than they were for absolute errors and the adjusted R^2 values were very small. Both of these results suggest that population size had little impact on bias; we return to this point later in the paper. As hypothesized, the growth rate had a significant positive effect on algebraic errors for both forecast horizons. Again, increasing the length of the forecast horizon had little impact on regression coefficients and adjusted R^2 values.

Complex Single-Variable Models

Although the simple single-variable models produced statistically significant results, neither variable was able to explain much of the county-to-county variation in forecast accuracy. Can more complex models improve on these results? To answer this question, we explored several alternative functional forms of the simple single-variable models. We refer to these as complex single-variable models because they include non-linear relationships and often include more than one term for each explanatory variable.

Our investigation of complex models was guided by the literature on population forecast accuracy. Several studies have found the relationship between population size and precision to weaken (or disappear completely) once a certain size has been reached (e.g., Smith 1987; Smith and Shahidullah 1995; Tayman 1996; Tayman, Schafer, and Carter 1998); this suggests that asymptotic functions such as the natural log or inverse might be applicable. Numerous studies have reported no consistent relationship between population size and bias (e.g., Isserman 1977; Rayer 2008; Smith 1987; Tayman 1996). Although this does not suggest any particular functional form, the natural log will reduce the impact of several very large counties. The positive relationship often found between growth rates and absolute percent errors suggests that a continuously increasing function, such as the natural log, might be applicable. Algebraic percent errors have been found to be large and negative for areas with large population declines and to become smaller but still negative as those declines become smaller, eventually becoming small but positive for areas with slowly growing populations and large and positive for areas with rapidly growing populations (Isserman 1977; Murdock et al 1984; Smith 1987; Tayman 1996). This suggests a polynomial function might be appropriate.

We considered several alternative models for each explanatory variable. We started with a simple linear model and sequentially added squared and cubed terms; we followed the same process using the natural log of each variable. We also considered inverse, compound, and power functions. Then, we selected the optimal model for each variable, defined as the model having the fewest parameters, simplest specification, and highest adjusted R^2 value. Our selection criteria were that an additional parameter or more complex specification (e.g., natural log rather than the variable itself) had to be statistically significant and add at least 1% to the adjusted R^2 value of the model. Because the discriminatory power of significance tests tends to

decline as sample size increases (Henkel 1976), these criteria helped us determine whether a particular parameter made a substantive contribution to the explanation of forecast error. Details of the model selection procedures are available from the authors on request.

(Table 4 about here)

The variables included in the optimal complex models, along with regression coefficients and adjusted R^2 values, are shown in Table 4. The top panel shows the results for absolute percent errors. For population size, the optimal model included the natural log and its square. Both terms had significant effects on forecast errors for both horizons. The natural log had a negative effect and its square had a positive effect, indicating that increases in population size reduced absolute percent errors at a declining rate. These results are consistent with our hypotheses and the findings of previous studies. Coefficients for both terms increased (in absolute value) as the forecast horizon became longer. Adjusted R^2 values were substantially larger than for the simple single-variable model and were larger for 10-year horizons than 20-year horizons. The asymptotic nature of the relationship between population size and absolute percent error can be seen in Figure 1, which shows that gains in precision became fairly small after counties reached a size of about 25,000.

(Figure 1 about here)

For growth rate, the optimal model included the natural log, its square, and its cube for 10-year horizons but only the natural log and its square for 20-year horizons. The coefficient for the natural log was negative for 10-year horizons, positive for 20-year horizons, and statistically significant for both. The squared term was positive and significant for both horizons and the cubed term was positive and significant for the 10-year horizon. The net result was the expected positive relationship between the absolute value of the growth rate and absolute percent errors.

This relationship was also asymptotic, especially for 20-year horizons (see Figure 2). Again, adjusted R^2 values were substantially larger than for the simple single-variable model and were larger for 10-year horizons than 20-year horizons.

(Figure 2 about here)

The bottom panel of Table 4 shows the results for algebraic percent errors. The optimal model for population size included the natural log but not its square; this implies that the flattening effect of the squared term found for absolute percent errors was not found for algebraic percent errors. This variable had a small but significant positive effect on algebraic errors for both horizons (see Figure 3). The slope of the curve was relatively flat (except for very small counties) and the low adjusted R^2 values indicate that population size did not explain much of the variation in algebraic percent errors. The regression coefficient increased with the length of the forecast horizon, but the adjusted R^2 value remained virtually unchanged.

(Figure 3 about here)

The optimal model for growth rate included linear, squared, and cubed terms for both forecast horizons. All three were significant for both horizons, with the linear and cubed terms having positive signs and the squared term a negative sign. The overall impact of the growth rate on algebraic percent errors was positive, but the size of the squared and cubed terms was so small that the relationship between the two was nearly linear (see Figure 4). Forecasts had the greatest downward bias in counties with the largest percentage declines, but the magnitude of the downward bias declined as growth rates increased, becoming positive when they reached approximately 10%. Regression coefficients for all three terms increased (in absolute value) with the length of horizon, as did the adjusted R^2 value.

(Figure 4 about here)

12

For both forecast horizons and both error measures, adjusted R^2 values were larger for growth rate models than population size models. This implies that differences in growth rates did a better job of explaining variations in forecast errors than did differences in population size. The particularly low R^2 values for population size in the algebraic error regressions once again demonstrate the lack of a consistent relationship between population size and the direction of forecast errors. For growth rates, adjusted R^2 values were roughly the same in regressions involving absolute percent errors as in regressions involving algebraic percent errors, suggesting that growth rates did about equally well in explaining variations in precision and bias.

Multivariate Models

Complex single-variable models clearly outperformed simple single-variable models in terms of discriminatory power. Can we raise discriminatory power even more by combining population size and growth rate in a multivariate model? What happens if we sequentially add prior error, census division, and launch year as explanatory variables? To answer these questions, we constructed a series of multivariate regression models. Model 1 included population size and growth rate as explanatory variables, using the optimal functional forms shown in Table 4. Model 2 added prior error to the variables in Model 1.[4] Model 3 added census divisions to the variables in Model 2 and Model 4 replaced census divisions with launch years. Model 5 contained all five explanatory variables.[5]

Basic regression results for the multivariate models are shown in Tables 5-8; a more complete description of model diagnostics is presented in the Appendix. Table 5 shows the results for absolute percent errors for 10-year horizons. For Model 1, coefficients for all the population size and growth rate terms were statistically significant and had the same signs as in

the optimal complex single-variable models. The adjusted R^2 value for the multivariate model was substantially higher than those found for either of the single-variable models, indicating that the multivariate model did a much better job of explaining variations in precision than did either of the single-variable models.

(Table 5 about here)

Results for population size and growth rate were consistent across all five models. In every instance, coefficients had the same sign and were statistically significant. The coefficients themselves were similar in every model (especially Models 2-5). This high level of consistency is striking and supports the validity of the results. Even as other explanatory variables were added, the effects of population size and growth rate remained largely unchanged.

As expected, prior absolute error had a significant positive effect on absolute percent errors: the greater the error during the base period, the larger the error over the forecast horizon. However, as indicated by the negative sign of the squared term, this effect declined as prior error became larger. The coefficients for both prior error terms changed very little from one model to the next.

The New England and the West North Central divisions had significant negative effects on absolute percent errors and the East South Central, West South Central, Mountain, and Pacific divisions had significant positive effects. Apparently, the first two divisions had characteristics making it easier to forecast the population precisely (compared to the reference group), whereas the latter four had characteristics making it more difficult. Coefficients were not statistically significant for the Mid Atlantic and East North Central divisions.[6]

Four of the first five launch years had significant positive effects on absolute percent errors while the last two had negative but not statistically significant effects. It appears that it

has become somewhat easier to construct precise population forecasts in recent decades. Based on the adjusted R^2 values, census division and launch year added about equally to the discriminatory power of the models.

Multivariate models clearly did a better job than single-variable models in explaining variations in the precision of population forecasts. Furthermore, adjusted R^2 values rose steadily as explanatory variables were added. This suggests that searching for other determinants of forecast error may lead to further improvements in the discriminatory power of the models.

Table 6 shows the results for absolute percent errors for 20-year horizons. For every model, the signs and levels of significance for the population size variables were identical to those found for 10-year horizons. For growth rate, however, the optimal model included only two terms rather than three. The sign of the linear term changed from negative to positive and the coefficient of the squared term remained positive but increased substantially in size. The overall impact of the growth rate on precision thus remained positive, just as it was for 10-year horizons. In most instances, the coefficients for the population size and growth rate terms were considerably larger for 20-year horizons than for 10-year horizons, suggesting that the effects of both variables persisted over time.

Prior error had the same signs and levels of significance as for 10-year horizons, but the coefficients themselves were substantially smaller, suggesting that the impact of this variable faded over time. The signs and levels of significance for most census divisions were the same for both 10- and 20-year horizons, but results for individual launch years were inconsistent from one horizon to another.

(Table 6 about here)

Table 7 shows the results for algebraic percent errors for 10-year horizons. For population size, coefficients were positive and statistically significant for four of the five models. The impact of population size was trivial, however: The adjusted R^2 value for the multivariate model containing population size and growth rate was no greater than for the complex model containing only the growth rate (see Table 4). For growth rate, coefficients for all three terms were statistically significant in every model and had the same signs as in the complex single-variable model. The growth rate coefficients themselves varied within a fairly narrow range, especially for Models 2-5. Again, we believe this supports the validity of the results.

(Table 7 about here)

The linear term for prior algebraic percent error had a significant negative effect on algebraic percent errors in Models 2-5 and the squared term had a significant but much smaller positive effect, yielding a negative overall effect, as hypothesized. Three census divisions had significant positive effects in Model 3 and four in Model 5; two had significant negative effects in both models. The first four launch years had positive errors and two of the last three had negative errors (compared to the reference group); this may reflect a shift in bias over time, but the evidence is weak. Adjusted R^2 values increased steadily as explanatory variables were added to the model, but the addition of the prior error and launch year variables had a much greater impact than did the addition of census divisions.

Table 8 shows the results for algebraic percent errors for 20-year horizons. In almost every instance, signs and levels of significance for the population size and growth rate variables were the same as for 10-year horizons. The coefficients for the growth rate variables were substantially larger for 20-year horizons than 10-year horizons, reflecting the persistence of those effects over time. For prior error, signs and levels of significance for the linear term were the

same as for 10-year horizons but the coefficients themselves were considerably smaller (there was no quadratic term for 20-year horizons). Moreover, whereas adding prior error to the first multivariate model substantially raised adjusted R^2 values for 10-year horizons, there were virtually no improvements for 20-year horizons. Again, we believe this suggests that the impact of prior error fades over time. The signs and levels of significance for census division and launch year were generally the same as for 10-year horizons, but the coefficients themselves were substantially larger.

(Table 8 about here)

In most instances, adjusted R^2 values for both 10- and 20-year forecast horizons were larger for regressions involving absolute percent errors than regressions involving algebraic percent errors. As noted previously, this suggests that the explanatory variables did a better job of explaining variations in precision than variations in bias. However, the addition of the launch year variables reduced this gap considerably, indicating that these variables had a greater impact on bias than on precision.

For absolute percent errors, adjusted R^2 values were larger for 10-year horizons than 20-year horizons for all five models. For algebraic percent errors, the same was true for every model except Model 1. This suggests that the ability of the explanatory variables to explain differences in forecast errors declines as the forecast horizon becomes longer. This is not surprising, of course: the longer the time period, the greater the likelihood that statistical relationships will change.

Relative Impact of Explanatory Variables

Our analysis thus far has focused on modeling the effects of population size and growth rate on forecast accuracy and evaluating the impact of adding other explanatory variables to the regression model. All the variables we considered were found to have statistically significant effects in most instances. We turn now to the question of which explanatory variable has the greatest impact on precision and bias. One way to answer this question is to measure the reduction in the adjusted R^2 value that occurs when one variable (including the complex forms of size, growth rate, and prior error variables) is removed from the fully specified multivariate model (Model 5). We interpret this reduction as a measure of each variable's contribution to the model's discriminatory power: the greater the reduction, the greater the impact of that variable on forecast error. The results are shown in Table 9.

(Table 9 about here)

For absolute percent errors, differences in growth rates contributed the most for both forecast horizons. Removing the growth rate variables reduced adjusted R^2 values by 6.5 and 8.9 percentage points for 10- and 20-year horizons, respectively. Population size was also important, with the third largest impact for both 10- and 20-year horizons (3.1 and 3.3 percentage points, respectively). Prior error was the second most important variable for 10-year horizons but the least important for 20-year horizons. Neither census division nor launch year had much impact for 10-year horizons, but launch year had the second largest impact for 20-year horizons.

For algebraic percent errors, differences in launch year had the greatest impact on adjusted R^2 values for 10-year horizons, followed closely by prior error. Removing these variables reduced adjusted R^2 values by 6.8 and 6.0 percentage points, respectively. Launch year remained important for 20-year horizons (reduction of 5.3 percentage points), but prior error did

not (reduction of only 0.2 percentage points). Growth rate had the third largest impact for 10-year horizons (reduction of 3.8 percentage points) and by far the greatest impact for 20-year horizons (reduction of 12.2 percentage points). Census division had relatively little impact on algebraic percent errors for either horizon and population size had virtually no impact at all.

What conclusions can we draw from these results? First, it is clear that—of the five explanatory variables we examined—the growth rate had the greatest impact on the precision of population forecasts. It was an important determinant of bias as well, especially for longer horizons.   Second, population size had a substantial impact on precision but not on bias. Although the regression analyses showed population size to have a statistically significant effect on algebraic percent errors, the magnitude of that effect was very small. Third, prior error had important short-term effects on both precision and bias, but its influence faded substantially over time. Fourth, differences in census division had significant but rather small effects on both precision and bias. Finally, differences in launch year had a greater effect on algebraic percent errors than on absolute percent errors, especially for 10-year horizons. As has been noted before (e.g., Smith and Sincich 1988), it appears that the direction of forecast errors varies more from one launch year to another than does the absolute size of those errors.

SUMMARY AND CONCLUSIONS

In this study, we used regression analysis to examine the effects of several explanatory variables on the precision and bias of population forecasts for a large sample of counties in the United States. Regression models make it possible to conduct more detailed analyses than can be done using more traditional approaches. Perhaps more important, they make it possible to test

hypotheses regarding the determinants of forecast accuracy. We believe regression analysis provides a powerful but under-utilized tool for evaluating population forecast errors.

We started by focusing on population size and growth rate, the variables most frequently considered in analyses of population forecast accuracy. Using data for individual counties, we developed several regression models, some using a simple linear form of a single explanatory variable, some using more complex forms of each variable, and some using both variables in a multivariate model. For simple single-variable models, we found that both population size and growth rate had statistically significant effects on precision and bias, but that neither variable could explain much of the variation in forecast errors. More complex forms of the single-variable models performed considerably better in this regard and a multivariate model containing both size and growth rate performed better still. Adding several other explanatory variables further enhanced the model's discriminatory power.

In terms of specific empirical results, we found that differences in population size had a significant non-linear impact on precision but little effect on bias and that differences in growth rates had significant effects on both precision and bias. These results are consistent with those reported in many studies using aggregate data. We also found that the explanatory variables did a better job explaining differences in precision than differences in bias and that growth rates had an asymptotic relationship with precision. To our knowledge, the latter two results have not been reported before.

We investigated the impact of three explanatory variables not usually considered in analyses of population forecast accuracy: prior error, census division, and launch year. We found that prior error had a statistically significant effect on both precision and bias, but that its impact declined considerably as the forecast horizon became longer. Differences in census

division and launch year also had significant effects on forecast accuracy, with the latter having a particularly large impact on bias. These two variables pick up the effects of omitted variables that influence forecast accuracy and vary across geographic areas and over time, respectively.[7]

This study provides perhaps the most comprehensive analysis yet of the determinants of population forecast accuracy and illustrates the value of developing regression models based on data for individual places. The fully specified multivariate models were able to explain between 24% and 33% of the variation in county forecast errors; this degree of discriminatory power is impressive for large cross-sectional data sets. Future research will undoubtedly extend our analysis to include additional explanatory variables and functional forms. Investigations of the impact of variables measuring economic, demographic, and environmental characteristics—and changes in those characteristics over time—are likely to be particularly important. Such research not only will add to our understanding of the determinants of population forecast accuracy, but is likely to lead to the development of more accurate forecasting models as well.

END NOTES

1. Counties include county equivalents in the District of Columbia, Louisiana, Missouri, and Nevada.

2. Several studies have found that base periods of 10 years are generally sufficient to produce the most accurate forecasts possible (e.g., Rayer 2008; Smith and Sincich 1990). We replicated our analyses using forecasts derived from 20-year base periods and found results similar to those reported here.

3. We show errors only for launch years 1920 through 1990 because those were the years used in the regression analyses.

4. We also investigated alternative functional forms of the prior error variable. A quadratic model outperformed a simple linear model for 10-year horizons but not for 20-year horizons. We therefore included both linear and squared terms in models covering 10-year horizons but not in models covering 20-year horizons. Logged models did not perform better than non-logged models for either horizon.

5. In order to include prior error as an explanatory variable, the regression analyses began with launch year 1920 for forecasts covering 10-year horizons and 1930 for forecasts covering 20-year horizons.

6. We investigated the effect of using states rather than census divisions as a measure of geographic differences. We found this change to slightly raise adjusted R^2 values (from 0.329 to 0.333 for APEs and from 0.299 to 0.310 for ALPEs) but to have virtually no impact on the regression coefficients of the other explanatory variables.

7. We also analyzed the relationship between forecast errors and population size and growth rate using data for individual census divisions and launch years. In most instances the results were similar to those reported here for the entire sample. The only exception was the relationship between algebraic percent errors and population size for individual launch years. Errors rose with population size for some launch years and declined for others. Once again, this reflects the absence of a consistent relationship between population size and bias.

REFERENCES

Belsley, D., E. Kuh, & R. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Chatterjee, S. and A. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons.

Chi, G. 2009. "Can Knowledge Improve Population Forecasts at Subcounty Levels?" *Demography* 46: 405-427.

Draper, N. and H. Smith. 1981. *Applied Regression Analysis, Second Edition*. New York: John Wiley & Sons.

Henkel, R. 1976. *Tests of Significance*. Beverly Hills: Sage.

Isserman, A. 1977. "The Accuracy of Population Projections for Subcounty Areas." *Journal of the American Institute of Planners* 43: 247-259.

Keyfitz, N. 1981. "The Limits of Population Forecasting." *Population and Development Review* 7: 579-593.

Lenze, D. G. 2000. "Forecast Accuracy and Efficiency: An Evaluation of Ex Ante Substate Long-Term Forecasts." *International Regional Science Review* 23: 201-226.

Long, J. F. 1995. "Complexity, Accuracy, and Utility of Official Population Projections." *Mathematical Population Studies* 5: 203-216.

Murdock, S., F. Leistritz, R. Hamm, S. Hwang, and B. Parpia. 1984. An Assessment of the Accuracy of a Regional Economic-Demographic Projection Model." *Demography* 21: 383-404.

Rayer, S. 2008. "Population Forecast Errors: A Primer for Planners." *Journal of Planning Education and Research* 27: 417-430.

Smith, S. K. 1987. "Tests of Forecast Accuracy and Bias for County Population Projections."

    Journal of the American Statistical Association 82: 991-1003.

Smith, S. K. and M. Shahidullah. 1995. "An Evaluation of Population Projection Errors for

    Census Tracts." *Journal of the American Statistical Association* 90: 64-71-508.

Smith, S. K. and T. Sincich. 1988. "Stability over Time in the Distribution of Population

    Forecasts Errors." *Demography* 25: 461-474.

Smith, S. K. and T. Sincich. 1990. "The Relationship between the Length of the Base Period and

    Population Forecast Errors." *Journal of the American Statistical Association* 85: 367-375.

Smith, S. K. and T. Sincich. 1992. "Evaluating the Forecast Accuracy and Bias of Alternative

    Population Projections for States." *International Journal of Forecasting* 8: 495-508.

Smith, S. K., J. Tayman. and D. Swanson. 2001. *State and Local Population Projections:*

    Methodology and Analysis. New York: Kluwer Academic/Plenum Publishers.

Stoto, M. 1983. "The accuracy of Population Projections." *Journal of the American Statistical*

    Association 78: 13-20.

Tayman, J. 1996. "The Accuracy of Small Area Population Forecasts Based on a Spatial

    Interaction Land Use Modeling System." *Journal of the American Planning Association*

    62: 85-98.

Tayman, J., E. Schafer, and L. Carter. 1998. "The Role of Population Size in the Determination

    and Prediction of Population Forecast Errors: An Evaluation Using Confidence Interval for

    Subcounty Areas." *Population Research and Policy Review* 17(1): 1-20.

White, H. R. 1954. "Empirical Study of the Accuracy of Selected Methods of Projecting State

    Population." *Journal of the American Statistical Association* 49: 480-498.

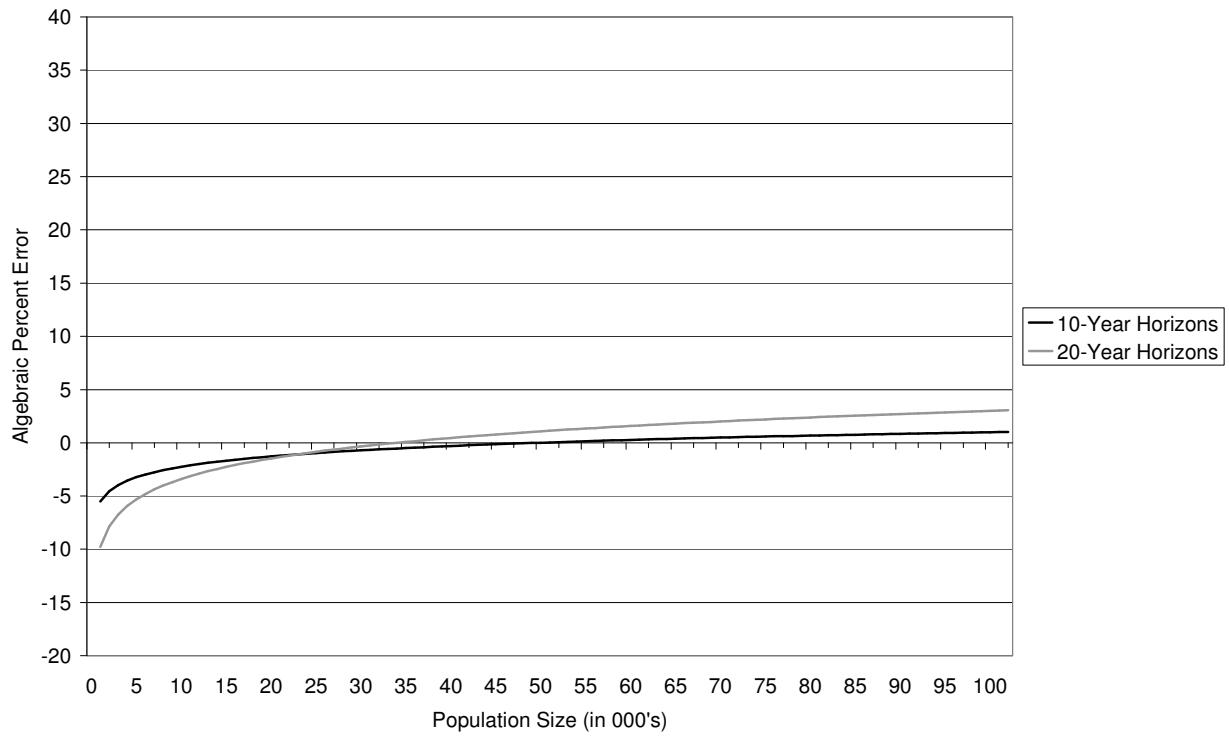**Figure 1. Prediction of Absolute Percent Errors Using Population Size[a]**
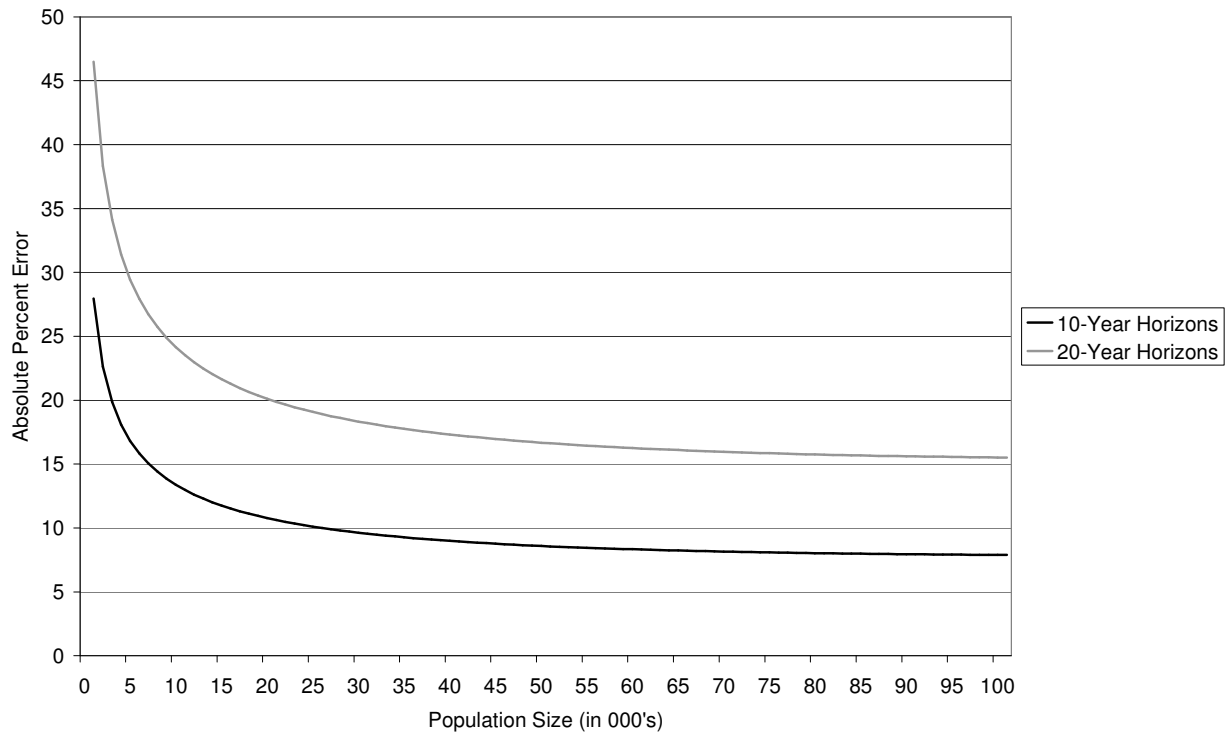
[a] Based on the quadratic function of the natural log of population size. For ease of interpretation, population size is expressed as untransformed values.

**Figure 2. Prediction of Absolute Percent Errors Using Growth Rate[a]**
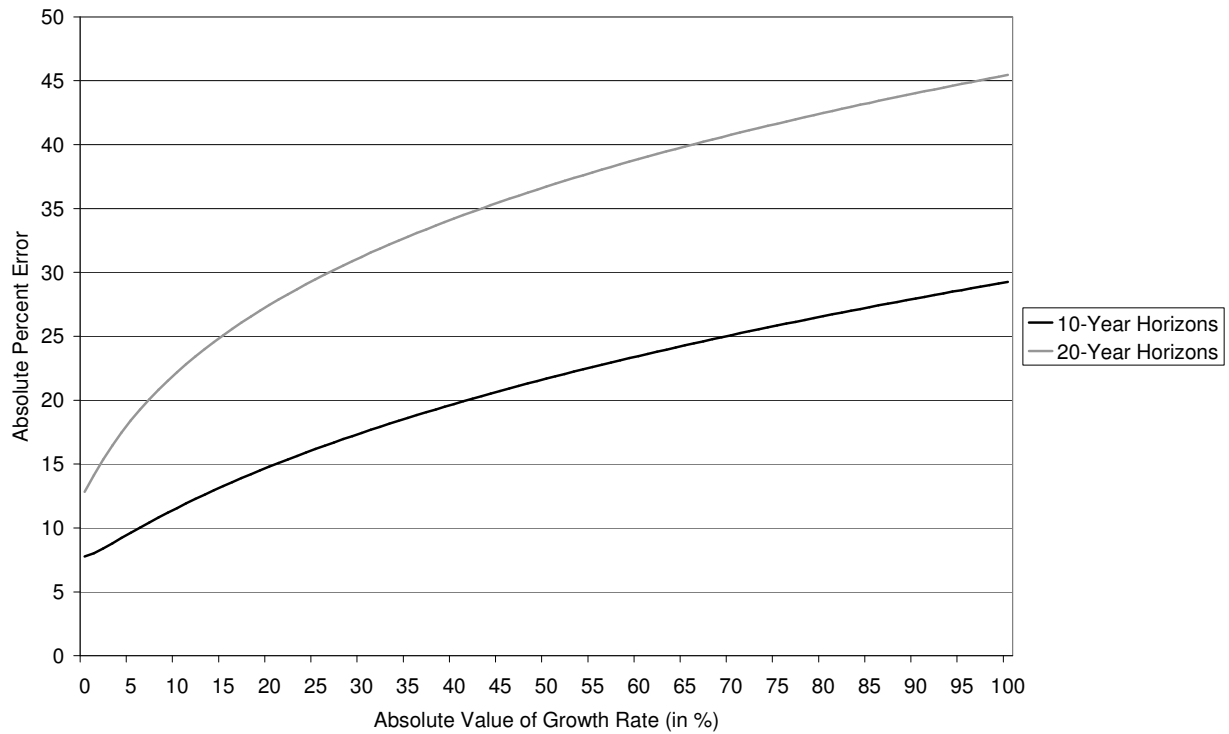
[a] Based on the cubic function of the natural log of the absolute value of the growth rate for 10-year horizons, and on the quadratic function of the natural log of the absolute value of the growth rate for 20-year horizons. For ease of interpretation, the horizontal axis is expressed as the untransformed absolute value of the growth rate.

Figure 3. Prediction of Algebraic Percent Errors Using Population Size[a]

[a] Based on the natural log of population size. For ease of interpretation, population size is expressed as untransformed values.

Figure 4. Prediction of Algebraic Percent Errors Using Growth Rate[a]
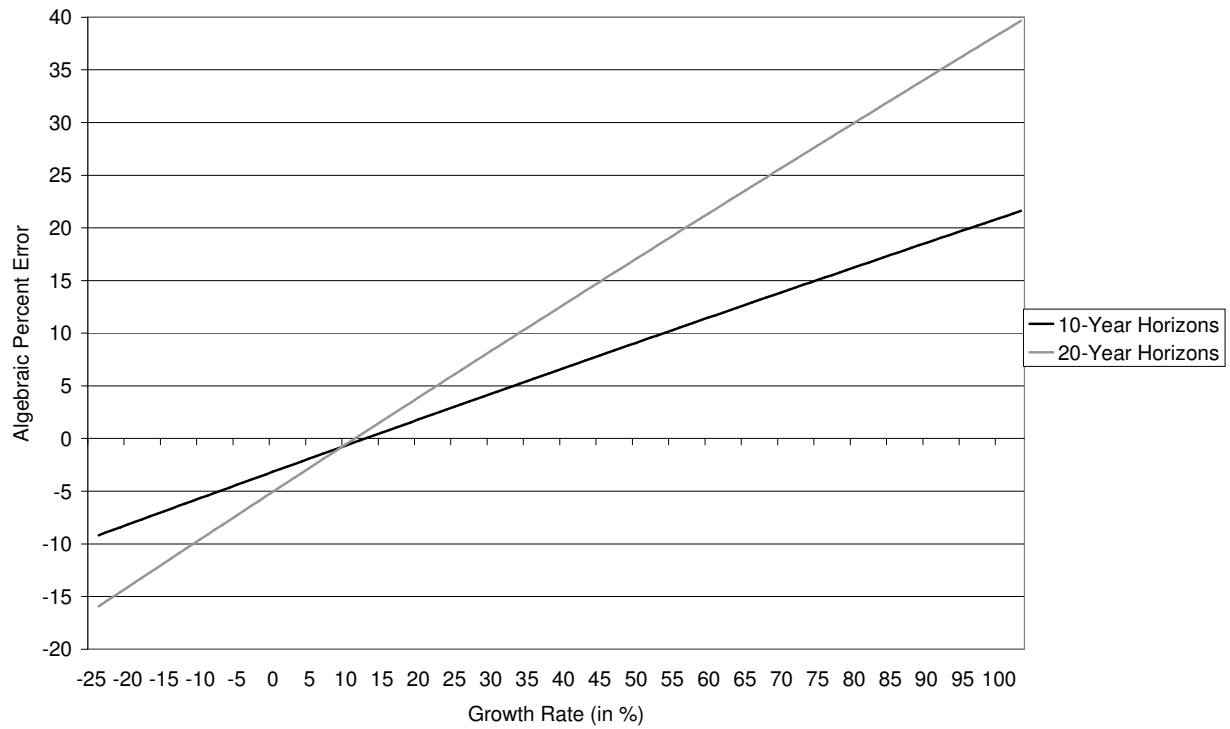
^{a} Based on the cubic function of the growth rate.

Table 1. County Population Size and Growth Rate Characteristics, 1900-2000

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Size											
Mean	26,126	30,787	35,044	40,615	43,336	49,453	58,502	66,122	72,950	79,054	88,574
Median	16,930	17,975	18,462	18,570	19,285	19,269	19,236	19,454	22,651	23,376	25,936
10th Percentile	4,104	5,530	5,877	6,514	6,417	6,151	5,786	5,574	6,066	5,827	6,001
90th Percentile	44,308	49,178	56,920	65,812	72,544	83,852	99,382	115,342	136,608	148,605	167,474
Growth Rate[a]											
Mean		44.0	7.7	20.8	5.9	4.0	5.3	5.5	15.4	3.6	10.7
Median		8.8	3.0	2.2	4.3	0.1	0.0	2.2	11.5	1.0	8.0
10th Percentile		-8.1	-10.8	-11.6	-8.8	-16.3	-16.4	-12.6	-4.0	-11.4	-4.0
90th Percentile		68.4	29.8	33.0	19.5	26.4	30.3	26.2	37.0	21.1	28.7
% Negative		29.6	40.2	42.6	30.7	49.6	50.0	43.0	17.8	46.7	22.2

[a] Percentage change over previous 10 years.

Table 2. Error Characteristics of Counties,
**   10-Year Forecast Horizons**

	MAPE	MALPE	Sample Size
Size			
< 5,000	20.9	-6.0	1,469
5,000 to 9,999	14.4	-2.5	2,966
10,000 to 24,999	11.3	-0.9	7,486
25,000 to 49,999	9.5	-0.5	4,135
50,000 to 99,999	9.3	0.3	1,948
100,000+	8.7	0.8	1,852
Growth Rate[a]			
< -10%	15.9	-14.7	2,799
-10.0 to 0%	8.9	-4.0	5,159
0.0 to 9.9%	9.0	0.1	5,654
10.0 to 24.9%	11.5	4.0	3,807
25.0 to 49.9%	15.6	6.8	1,704
50.0+%	27.4	15.4	733
Prior Abs. % Error			
< 2.0	8.5	-1.4	2,499
2.0 to 3.9	8.7	-1.3	2,433
4.0 to 7.9	9.2	-0.9	4,300
8.0 to 14.9	10.5	-1.1	5,142
15.0 to 24.9	13.6	-0.6	3,251
25.0+	23.2	-2.1	2,231
Prior Alg. % Error			
<-15.0	19.4	9.1	2,646
-15.0 to -8.0	11.4	2.9	2,643
-7.9 to 0.0	9.0	0.0	4,746
0.0 to 7.9	8.7	-2.4	4,486
8.0 to 14.9	9.6	-5.2	2,499
15.0+	15.7	-10.8	2,836

[a] Percentage change for 10 years prior to launch year

Table 2 (con). Error Characteristics of Counties, 10-Year Forecast Horizons

	MAPE	MALPE	Sample Size
Census Division			
New England	6.6	-0.8	456
Mid Atlantic	7.3	-0.5	1,152
East North Central	8.6	-1.1	3,432
West North Central	9.9	-0.2	4,512
South Atlantic	10.8	-1.9	3,000
East South Central	11.5	-1.4	2,632
West South Central	17.3	-1.0	2,872
Mountain	21.7	-2.7	880
Pacific	15.6	-2.9	920
Launch Year			
1920	13.2	-0.6	2,482
1930	14.1	0.8	2,482
1940	13.2	3.0	2,482
1950	9.7	-2.1	2,482
1960	10.6	-2.4	2,482
1970	12.5	-9.2	2,482
1980	10.9	8.7	2,482
1990	9.0	-7.3	2,482

Table 3. Simple Single-Variable Regression Models:
Unstandardized Coefficients and Adjusted R^2 Values

	Horizon Length	
	10-year	20-year
Absolute % Errors		
Population Size		
Size	-0.003 ***	-4.5E-06 ***
Adjusted R^2	0.003	0.002
Growth Rate		
GR-Abs	0.023 ***	0.026 ***
Adjusted R^2	0.031	0.017
Algebraic % Errors		
Population Size		
Size	2.9E-06 ***	7.8E-06 ***
Adjusted R^2	0.001	0.003
Growth Rate		
GR-Alg	0.032 ***	0.043 ***
Adjusted R^2	0.032	0.024

*** Significant at the .001
 ** Significant at the .01
  * Significant at the .05

Table 4. Complex Single-Variable Regression Models:
Unstandardized Coefficients and Adjusted R^2 Values

	Horizon Length	
	10-year	20-year
Absolute % Errors		
Population Size		
Ln Size	-20.150 ***	-30.493 ***
(Ln Size)^2	0.857 ***	1.290 ***
Adjusted R^2	0.091	0.065
Growth Rate		
Ln GR-Abs	-0.514 ***	0.673 **
Ln (GR-Abs)^2	0.451 ***	1.434 ***
Ln (GR-Abs)^3	0.143 ***	---
Adjusted R^2	0.153	0.126
Algebraic % Errors		
Population Size		
Ln Size	1.413 ***	2.774 ***
Adjusted R^2	0.010	0.011
Growth Rate		
GR-Alg	0.249 ***	0.449 ***
(GR-Alg)^2	-8.3E-05 ***	-1.5E-04 ***
(GR-Alg)^3	6.8E-09 ***	1.2E-08 ***
Adjusted R^2	0.149	0.164

*** Significant at the .001
 ** Significant at the .01
  * Significant at the .05

Table 5. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, 10-Year Horizons (Absolute % Errors)

	Model 1	Model 2	Model 3	Model 4	Model 5
Ln Size	-14.778 ***	-9.220 ***	-8.764 ***	-9.751 ***	-9.366 ***
$(\text{Ln Size})^2$	0.582 ***	0.347 ***	0.329 ***	0.375 ***	0.362 ***
Ln GR-Abs	-0.613 ***	-0.400 ***	-0.416 ***	-0.305 **	-0.322 ***
$(\text{Ln GR-Abs})^2$	0.487 ***	0.390 ***	0.357 ***	0.408 ***	0.375 ***
$(\text{Ln GR-Abs})^3$	0.141 ***	0.101 ***	0.098 ***	0.097 ***	0.094 ***
Prior-Abs		0.231 ***	0.209 ***	0.225 ***	0.203 ***
Prior-Abs^2		-1.3E-04 ***	-1.1E-04 ***	-1.2E-04 ***	-1.1E-04 ***
New England			-1.073 *		-1.187 *
Mid Atlantic			0.111		-0.016
East North Central			-0.170		-0.190
West North Central			-1.026 ***		-0.993 ***
East South Central			1.300 ***		1.333 ***
West South Central			2.811 ***		2.840 ***
Mountain			4.301 ***		4.309 ***
Pacific			1.385 ***		1.282 ***
1920				1.568 ***	1.669 ***
1930				2.800 ***	2.866 ***
1940				2.953 ***	2.964 ***
1950				-1.249 **	-1.197 ***
1970				2.386 ***	2.361 ***
1980				-0.343	-0.273
1990				-0.362	-0.429
Adjusted R^2	0.242	0.299	0.313	0.315	0.329

*** Significant at the .001

**   Significant at the .01

*   Significant at the .05

Table 6. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, 20-Year Horizons (Absolute % Errors)

	Model 1	Model 2	Model 3	Model 4	Model 5
Ln Size	-21.510 ***	-16.589 ***	-17.239 ***	-18.206 ***	-19.065 ***
(Ln Size)^2	0.823 ***	0.619 ***	0.660 ***	0.705 ***	0.758 ***
Ln GR-Abs	0.531 **	0.414 *	0.325	0.652 ***	0.560 **
(Ln GR-Abs)^2	1.512 ***	1.367 ***	1.284 ***	1.413 ***	1.333 ***
Prior-Abs		0.127 ***	0.100 ***	0.122 ***	0.096 ***
New England			-2.762 *		-2.942 **
Mid Atlantic			-0.774		-1.116
East North Central			-2.128 ***		-2.160 ***
West North Central			-3.310 ***		-3.155 ***
East South Central			2.392 ***		2.468 ***
West South Central			6.475 ***		6.371 ***
Mountain			4.442 ***		4.228 ***
Pacific			-0.844		-1.417
1930				-0.504	-0.502
1940				5.070 ***	4.801 ***
1950				-4.562 ***	-4.664 ***
1970				-4.274 ***	-4.573 ***
1980				-9.252 ***	-9.297 ***
Adjusted R^2	0.194	0.214	0.236	0.257	0.278

*** Significant at the .001
** Significant at the .01
* Significant at the .05

Table 7. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, 10-Year Horizons (Algebraic % Errors)

	Model 1	Model 2	Model 3	Model 4	Model 5
Ln Size	0.157	0.395 ***	0.589 ***	0.546 ***	0.772 ***
GR-Alg	0.247 ***	0.153 ***	0.167 ***	0.138 ***	0.151 ***
$(GR\text{-}Alg)^2$	-8.3E-05 ***	-5.2E-05 ***	-5.7E-05 ***	-4.7E-05 ***	-5.1E-05 ***
$(GR\text{-}Alg)^3$	6.8E-09 ***	4.3E-09 ***	4.7E-09 ***	3.9E-09 ***	4.3E-09 ***
Prior-Alg		-0.265 ***	-0.259 ***	-0.259 ***	-0.253 ***
$Prior\text{-}Alg^2$		1.3E-04 ***	1.3E-04 ***	1.3E-04 ***	1.3E-04 ***
New England			0.274		0.093
Mid Atlantic			0.655		0.398
East North Central			0.968		0.853 *
West North Central			3.812 ***		3.748 ***
East South Central			1.486 ***		1.411 ***
West South Central			1.330 ***		1.403 ***
Mountain			-1.411 *		-1.161 *
Pacific			-3.907 ***		-3.754 ***
1920				4.190 ***	4.144 ***
1930				2.913 ***	2.868 ***
1940				6.045 ***	6.034 ***
1950				1.820 ***	1.817 ***
1970				-6.918 ***	-6.928 ***
1980				7.809 ***	7.674 ***
1990				-1.993 ***	-2.071 ***
Adjusted R^2	0.149	0.221	0.232	0.289	0.299

*** Significant at the .001
**   Significant at the .01
*   Significant at the .05

Table 8. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, 20-Year Horizons (Algebraic % Errors)

	Model 1	Model 2	Model 3	Model 4	Model 5
Ln Size	0.367	0.365	0.728 ***	0.482 *	0.861 ***
GR-Alg	0.446 ***	0.428 ***	0.457 ***	0.406 ***	0.434 ***
$(GR-Alg)^2$	-1.5E-04 ***	-1.4E-04 ***	-1.6E-04 ***	-1.4E-04 ***	-1.5E-04 ***
$(GR-Alg)^3$	1.2E-08 ***	1.2E-08 ***	1.3E-08 ***	1.1E-08 ***	1.2E-08 ***
Prior-Alg		-0.043 ***	-0.041 ***	-0.040 ***	-0.039 ***
New England			-0.121		-0.225
Mid Atlantic			3.246 **		3.098 **
East North Central			2.885 ***		2.823 ***
West North Central			8.882 ***		8.746 ***
East South Central			6.138 ***		6.053 ***
West South Central			5.192 ***		5.297 ***
Mountain			-4.623 ***		-4.375 ***
Pacific			-8.924 ***		-8.582 ***
1930				14.179 ***	14.091 ***
1940				18.382 ***	18.389 ***
1950				7.132 ***	7.180 ***
1970				2.376 **	2.366 **
1980				16.551 ***	16.225 ***
Adjusted R^2	0.164	0.166	0.187	0.220	0.240

*** Significant at the .001
** Significant at the .01
* Significant at the .05

Table 9. Adjusted R^2 and Reduction in Adjusted R^2 after Removing Explanatory Variables from Model 5

	Absolute Percent Error		Algebraic Percent Error	
Horizon Length	10	20	10	20
Adjusted R^2 (Model 5)	0.329	0.278	0.299	0.240
Variable Removed [a]				
Population Size	0.031	0.033	0.002	0.001
Growth Rate	0.065	0.089	0.038	0.122
Prior Error	0.042	0.010	0.060	0.002
Census Division	0.014	0.022	0.010	0.020
Launch Year	0.016	0.042	0.068	0.053

[a] Includes all terms for each variable

APPENDIX
EVALUATION OF REGRESSION EQUATIONS

Diagnostic Tools

The following diagnostics provide a more complete evaluation of the regression equations. The variables are those presented in the main body of the paper. We use several diagnostic tools, as described in Belsley, Kuh, and Welsch (1980), Chatterjee and Hadi (1988), Draper and Smith (1981), and elsewhere. For simplicity, we focus primarily on Model 5.

We calculated Cook's D and DFBETA to evaluate the impact of influential observations on the regression results. Cook's D assesses the influence (i.e., scaled distance) of an observation on the estimated set of coefficients. Values exceeding the conventional cut-off point $(4 / n)$ indicate an observation that may excessively influence the regression results. The DFBETA diagnostic assesses the effect of an individual observation on each estimated parameter in the model; for each parameter estimate, it calculates for each observation the standardized difference in the parameter estimate due to deleting the observation. Absolute values exceeding the conventional cut-off point $(2 / \sqrt{n})$ indicate that a particular observation may be excessively influential. We also plotted residuals against predicted y-values to check for heteroscedasticity and model misspecification, and created normal probability plots to check the normality assumption.

The Variance Inflation Factor (VIF) is used to examine multicollinearity. The VIF expresses the strength of the association between any two explanatory variables in the model. A VIF value higher than 10 indicates that multicollinearity may be present. We also examined the Predicted REsidual Sums of Squares (PRESS) statistic as a measure of a model's predictive power. The PRESS statistic measures how well the model predicts observed responses by fitting

the model repeatedly, leaving out one observation each time. In each repetition, the model is used to predict the deleted observation. There are no precise rules for interpreting PRESS values, but smaller values indicate greater predictive power.

Cook's D and DFBETA

The tests show that Model 5 was affected by influential observations for both precision and bias for 10-year horizons. For precision, 4.8% of the 19,856 observations had a Cook's D value above the cut-off point, as did 5.1% of the observations for bias. For each explanatory variable, between 1.7% and 7.2% of the observations had a DFBETA value exceeding the cut-off point for precision; for bias, the range was 2.0% to 7.1%. For precision, 26.5% of the observations had at least one explanatory variable with a DFBETA value exceeding the cut-off point; for bias, the proportion was 26.2% (data not shown).

To examine the impact of influential observations, we reran Model 5 using only observations with Cook's D and DFBETA values below the cut-off points. Excluding the influential observations substantially raised the explanatory power of both equations. For 10-year horizons, adjusted R^2 values increased from 0.329 to 0.491 for the precision regression and from 0.299 to 0.600 for the bias regression (see Table A1). Coefficients were generally similar to the model that included all observations with the following exceptions: for precision, the coefficient for Prior-Abs^2 became positive; for bias, the coefficient for Prior-Alg^2 lost its statistical significance. In addition, some dummy variable coefficients changed signs and significance levels (data not shown).

(Table A1 about here)

Results for 20-year horizons were generally similar to those for 10-year horizons. Excluding influential observations raised adjusted R^2 values from 0.278 to 0.470 for the

precision regression, and from 0.240 to 0.517 for the bias regression (see Table A1). For

precision, the coefficient for Ln Gr-Abs changed sign and lost its statistical significance; for bias,

the coefficient for Ln Size changed sign and lost its statistical significance.

Residual and Normal Probability Plots

Excluding influential observations had little impact on the shape of plots showing the

relationship between residuals and predicted y-values for precision and bias for 10- and 20-year

horizons (see Figure A1). However, the range was considerably narrower when influential

observations were excluded. None of the plots indicated the presence of model misspecification

or heteroscedasticity in the residuals.

(Figure A1 about here)

For the full data set, normal probability plots showed a slight S-shaped pattern indicative

of a non-normally distributed error term (data not shown). When influential observations were

excluded, the plots showed almost a perfectly straight line. We do not consider small deviations

from normality to be a problem. The central limit theorem says that when errors are not

normally distributed, a sufficiently large sample size will produce a normal sampling distribution

of the regression coefficients. Therefore, violations of this assumption usually have little or no

impact on substantive conclusions for large samples.

Variance Inflation Factor (VIF)

For both 10- and 20-year horizons, all independent variables in Model 5 except Ln Size and

$(Ln\ Size)^2$ had VIF values below the cut-off point in the precision regression; in the bias

regression, the same was true for all variables except GR-Alg, $(GR-Alg)^2$ and $(GR-Alg)^3$ (data

not shown). For precision, rerunning the model without the higher order term for population size

had only a trivial effect on the adjusted R^2 value; for bias, rerunning the model without the

higher order terms for growth rate resulted in a larger drop, especially for 20-year horizons (see

Table A2). All remaining coefficients for population size, growth rate, and prior error kept the

same signs, and all but Ln GR-Abs kept the same level of significance (data not shown). Thus,

although multicollinearity was present for the two population size variables for precision, and for

the three growth rate variables for bias, its presence had no substantive impact on the regression

results.

(Table A2 about here)

PRESS

For precision regressions with 10-year horizons, PRESS values for both population size and

growth rate were lower for complex models than for simple models and declined as explanatory

variables were added to the model (see Table A2). These lower values reflect greater predictive

power. However, rerunning Model 5 without Ln Size^2 (because of VIF issues) raised the PRESS

value slightly. Results were similar for 20-year horizons. Overall, PRESS values and adjusted

R^2 values were consistent with each other and both pointed to Model 5 as the preferred

specification.

For bias regressions with 10-year horizons, PRESS values and adjusted R^2 values were

also consistent with each other in most instances. PRESS values for population size were lower

for the complex model than for the simple model, whereas for growth rate the opposite was true.

PRESS values declined and adjusted R^2 values increased as explanatory variables were added to

Model 1; Model 5 had the highest adjusted R^2 value, but PRESS values were slightly higher than

for Model 4, which shows that the census division dummy variables had little impact on bias. Excluding the higher order terms for growth rate from Model 5 (because of VIF issues) produced the lowest PRESS value, but adjusted R^2 values dropped slightly compared to Models 4 and 5.

For bias regressions with 20-year horizons, however, PRESS values and adjusted R^2 values were often inconsistent with each other. The complex population size model had a slightly lower PRESS value than the simple population size model, but for the growth rate the complex model had a substantially higher PRESS value. The first result was consistent with the adjusted R^2 values, but the second was not. Rerunning Model 5 without the higher order terms for growth rate (again, because of VIF issues) produced the lowest PRESS values for both horizons, but adjusted R^2 values declined substantially. For bias regressions with 20-year horizons, then, PRESS values and adjusted R^2 values pointed in different directions with respect to the selection of the preferred model.

Summary

The diagnostics shown here largely support the empirical results presented in the main body of the paper. For precision, Model 5 is clearly the preferred model according to both adjusted R^2 and PRESS values; furthermore, the multicollinearity between the two population size variables and the influential observations detected by Cook's D and DFBETA statistics were found to have no substantive impact on the regression coefficients for any of the explanatory variables. We also found that excluding influential observations led to substantial increases in the model's explanatory power.

For bias, the findings were not quite as clear-cut. Both VIF and PRESS values suggested that there were statistical issues regarding the higher order terms of the growth rate variable. PRESS values and adjusted R^2 values did not always provide consistent results regarding the

choice of an optimal model, especially for 20-year horizons; this again illustrates the difficulty of explaining and predicting forecast bias. Although these results do not change the general conclusions reported in this study, they highlight the need for additional research on the use of regression models for evaluating the accuracy of population forecasts.

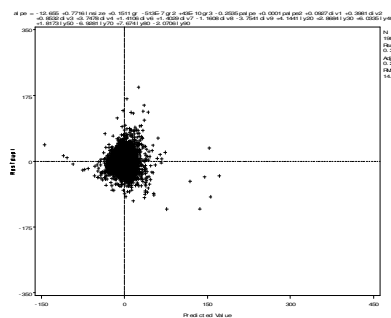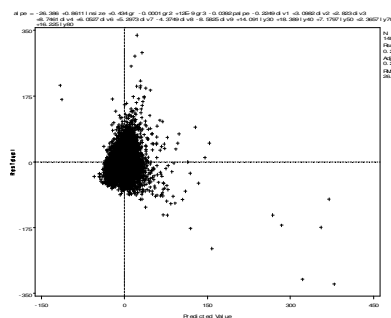**Figure A1. Residual vs. Predicted Values: All Observations:**

Excluding Influential Observations:

a. 10-Year Absolute

e. 10-Year Absolute
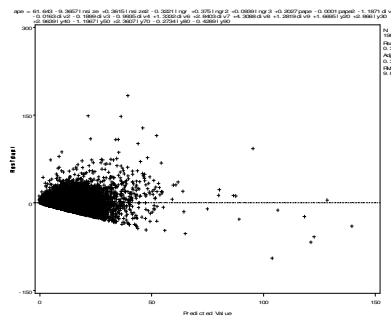
b. 20-Year Absolute
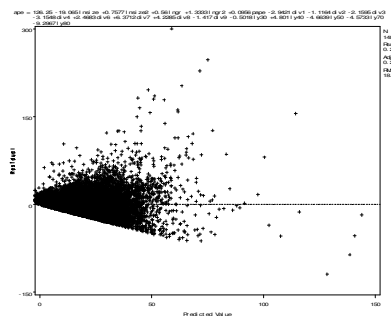
f. 20-Year Absolute

c. 10-Year Algebraic

g. 10-Year Algebraic
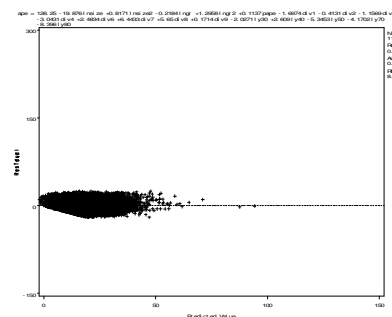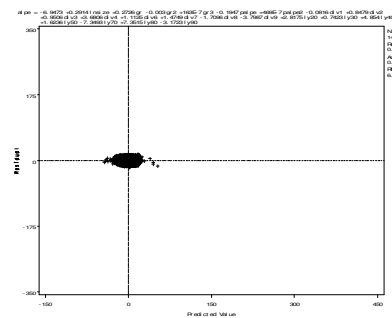

d. 20-Year Algebraic
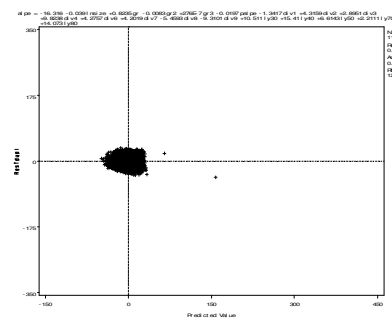

h. 20-Year Algebraic

46

Table A1a. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, Model 5 (Absolute % Errors)

	10 Year Horizons		20 Year Horizons	
	All	Reduced	All	Reduced
Ln Size	-9.366 ***	-8.543 ***	-19.065 ***	-19.876 ***
(Ln Size)^2	0.362 ***	0.333 ***	0.758 ***	0.817 ***
Ln GR-Abs	-0.322 ***	-0.463 ***	0.560 **	-0.218
(Ln GR-Abs)^2	0.375 ***	0.317 ***	1.333 ***	1.296 ***
(Ln GR-Abs)^3	0.094 ***	0.104 ***	-	-
Prior-Abs	0.203 ***	0.176 ***	0.096 ***	0.114 ***
Prior-Abs^2	-1.1E-04 ***	4.8E-04 ***	-	-
Adjusted R^2	0.329	0.491	0.278	0.470

Table A1b. Multivariate Regression Models: Unstandardized Coefficients and Adjusted R^2 Values, Model 5 (Algebraic % Errors)

	10 Year Horizons		20 Year Horizons	
	All	Reduced	All	Reduced
Ln Size	0.772 ***	0.291 ***	0.861 ***	-0.039
GR-Alg	0.151 ***	0.273 ***	0.434 ***	0.824 ***
(GR-Alg)^2	-5.1E-05 ***	-0.003 ***	-1.5E-04 ***	-0.008 ***
(GR-Alg)^3	4.3E-09 ***	1.6E-05 ***	1.2E-08 ***	2.7E-05 ***
Prior-Alg	-0.253 ***	-0.195 ***	-0.039 ***	-0.0197 ***
Prior-Alg^2	1.3E-04 ***	4.7E-05	-	-
Adjusted R^2	0.299	0.600	0.240	0.517

*** Significant at the .001

** Significant at the .01

* Significant at the .05

Note: "Reduced" equations are those excluding influential observations.

Table A2. Predicted Residual Sums of Squares and Adjusted R^2 Values

	10-Year			
	APE		ALPE	
Model	PRESS	Adj. R^2	PRESS	Adj. R^2
Pop. Size Simple	2,874,957	0.003	5,557,136	0.001
Growth Rate Simple	2,815,647	0.031	5,441,129	0.032
Pop. Size Complex	2,623,033	0.091	5,509,648	0.010
Growth Rate Complex	2,446,680	0.153	5,874,646	0.149
Model 1	2,192,807	0.242	5,886,284	0.149
Model 2	2,051,180	0.299	4,727,909	0.221
Model 3	2,002,676	0.313	4,806,092	0.232
Model 4	2,004,838	0.315	4,233,964	0.289
Model 5	1,956,223	0.329	4,286,873	0.299
Model 5 Excl. LnSize^2	1,966,905	0.324	-	-
Model 5 Excl. GR^2+GR^3	-	-	4,127,578	0.269

	20-Year			
	APE		ALPE	
Model	PRESS	Adj. R^2	PRESS	Adj. R^2
Pop. Size Simple	6,771,000	0.002	13,569,684	0.003
Growth Rate Simple	6,709,976	0.017	13,425,879	0.024
Pop. Size Complex	6,342,378	0.065	13,449,513	0.011
Growth Rate Complex	5,934,604	0.126	16,300,276	0.164
Model 1	5,470,183	0.194	16,373,708	0.164
Model 2	5,339,080	0.214	16,104,498	0.166
Model 3	5,192,699	0.236	16,902,840	0.187
Model 4	5,049,198	0.257	14,843,484	0.220
Model 5	4,906,521	0.278	15,576,921	0.240
Model 5 Excl. LnSize^2	4,950,197	0.272	-	-
Model 5 Excl. GR^2+GR^3	-	-	11,905,404	0.133