Go West [and South], young man:
Modeling population change within the modern United States

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Jacey Planteen

Department of Statistics and Analytical Sciences

Faculty Mentor: Dr. Brad Barney

ABSTRACT

This research modeled population change between 2000 and 2010 by county in the United States utilizing various demographic, social, and economic variables from decennial census data. A useful population change model could help communities ensure infrastructure keeps pace with a growing – or declining – population size. The population change model was developed using a variety of techniques, including ordinary least squares regression and robust regression. Census observations were split into training and test data sets using the DUPLEX algorithm in order to both develop and validate the model. Ultimately, the final model only explains approximately 58% of the overall variation in population change; clearly not all important factors for population growth and decline may be captured in the census data alone. However, numerous regressors are significant and could be useful indicators. Some of the strongest indicators are the percent of the population under five years old, the percent enrolled in college, the percent of the population in the labor force, geographic region, and the difference between the average family size and average household size. Thus, though this model has substantial error, it does have some utility and value for understanding population growth and decline.

INTRODUCTION

The chief objective of this research was to develop a model for county-level population change between 2000 and 2010 in the United States utilizing regression techniques along with demographic and socioeconomic information from decennial census data. A useful model for population change has obvious value: anticipating population growth and decline allows for proper infrastructure planning, governmental budgeting, and disaster planning. Existing population change models have a much more theoretical basis. For example, the U.S. Census Bureau publishes estimates and projections using a cohort-component model. This model estimates population by its individual components: births, deaths, and migration, using data from other sources, such as death certificates for mortality rates.[1] In contrast, the model developed by this research is novel in that:

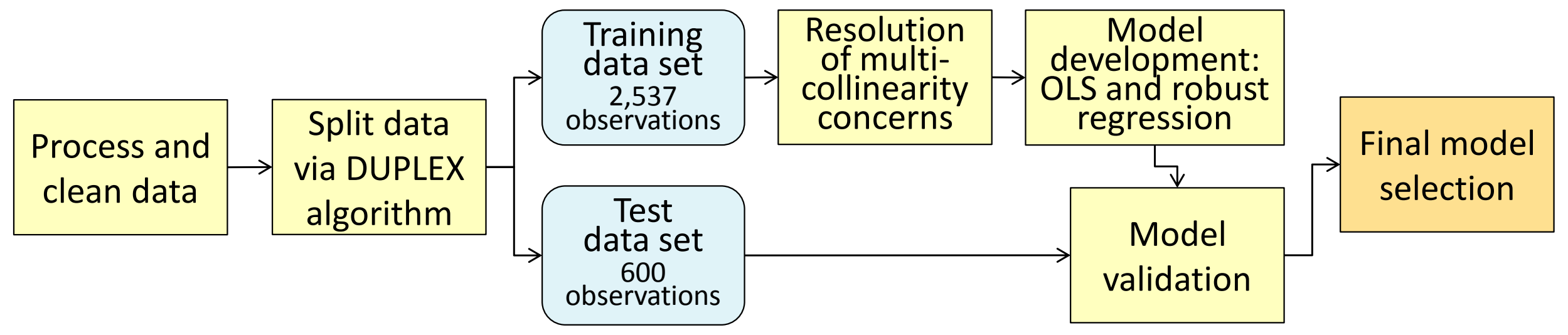
- Data from only one source, the decennial census, were used.
- Only overall population change and not its individual components were considered.
- A variety of additional factors – mostly of a socioeconomic nature – were included as possible indicators for population change.

Thus, though the main goal was to develop a useful regression model for population change, a secondary output was to identify potential indicators for population growth and decline.

[1] "Methodology for the United States Population Estimates: Vintage 2014: Nation, States, Counties, and Puerto Rico – April 1, 2010 to July 1, 2014," United States Census Bureau, accessed Nov 11, 2015, https://www.census.gov/popest/methodology/2014-natstcopr-meth.pdf.

METHODS

This analysis used data from the 2000 and 2010 decennial censuses, which were accessed from the American Fact Finder Census Download Center.[2] Only the county population totals were needed from the 2010 census, but a wide variety of demographic, social, and economic variables were utilized from the 2000 census. For the 2000 census, key demographic variables were collected at a 100% sampling rate. However, only one out of six households were asked to provide more detailed social and economic information.[3] Thus, it would expected for more error to exist in the social and economic variables due to the sampling method.



Process and clean data → Split data via DUPLEX algorithm → Training data set (2,537 observations) → Resolution of multi-collinearity concerns → Model development: OLS and robust regression → Model validation → Final model selection

Split data via DUPLEX algorithm → Test data set (600 observations) → Model validation

The basic methodology for model development is shown in the diagram above and additional details are discussed below. The DUPLEX algorithm was used to split the data to ensure the validation observations would properly test the model's predictive quality. Several tactics were taken to resolve multi-collinearity concerns. In some cases, highly-correlated variables could be removed from the model with no significant effect; in other cases, modified variables had to be created to reduce these concerns. The response variable originally chosen was the ratio of the 2010 population to the 2000 population; however, Box-Cox analysis indicated that the inverse had better properties and was thus used. In addition to ordinary least squares regression, robust regression using M-estimation with Tukey's bisquare function was used to limit the impact of outliers in the data. Simple first order models and higher order models using interactions of the geographic region with numeric variables were considered. A given model's predictive abilities weighed heavily in the selection of the final, "best" model. A significance level of 0.05 was used for all analyses.

[2] "Measuring America: the Decennial Censuses from 1970 to 2000," United States Census Bureau, prepared by J. Gauthier, issued 2002, accessed Nov 11, 2015, http://www2.census.gov/library/publications/decennial/pol_02-ma.pdf.

[3] Decennial Census Data 2000 and 2010 data sets, American Fact Finder Census Download Center, United States Census Bureau, accessed October 17, 2015, http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

Figure 1. Actual Population Change in the United States from 2000 to 2010

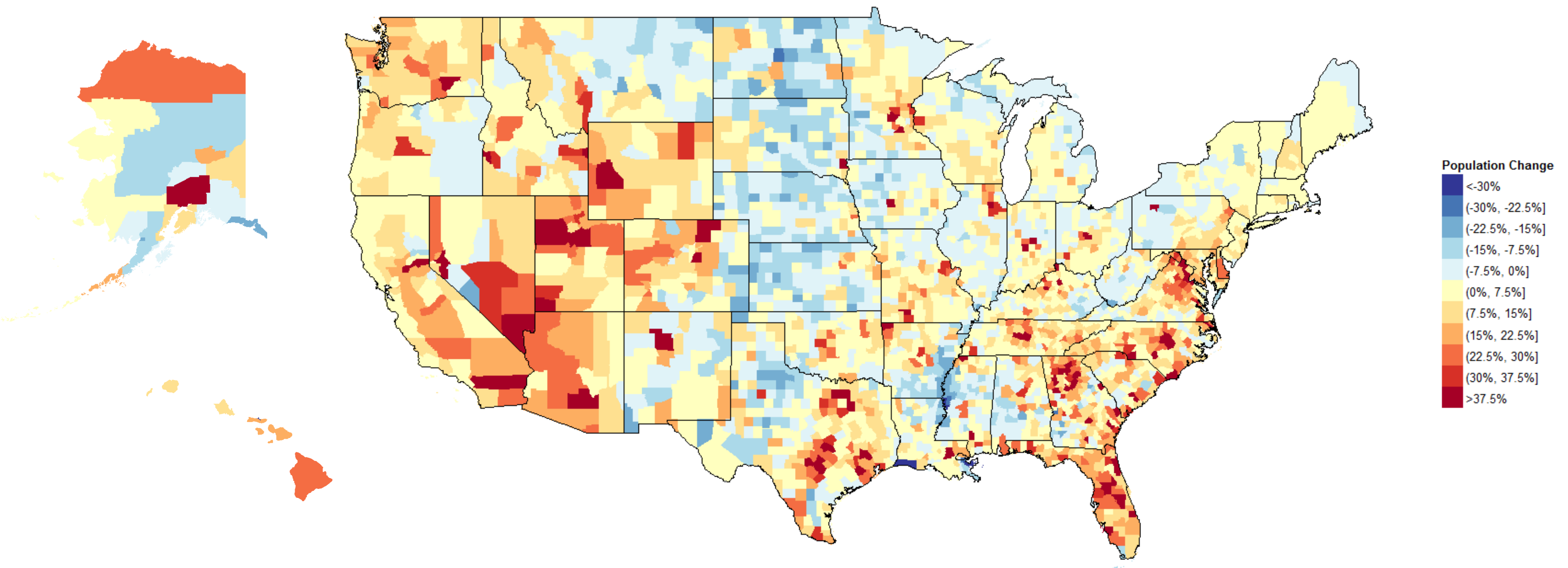


Table 1. Model Performance Comparisons

Model	Number of Terms	Model Development R^2	Model Development Std. Error	Model Validation R$^2_{predicted}$	Model Validation Std. Error
1st Order Ordinary Least Squares	27	0.581	0.0680	0.513	0.0900
1st Order Robust Regression: OLS Structure	25	0.571	0.0688	0.508	0.0903
1st Order Robust Regression	31	0.569	0.0691	0.487	0.0927
Ordinary Least Squares with Interactions	64	0.625	0.0659	0.512	0.0932
Robust Regression with Interactions: OLS Structure	64	0.613	0.0659	0.531	0.0913
Robust Regression with Interactions	69	0.598	0.0672	0.508	0.0940

Table 2. Final Model Coefficient Estimates

Parameter	Estimate	Stdized Estimate	t value	Pr(>|t|)	VIF
(Intercept)	1.546	0.000	10.8	< 0.001	
Region: Midwest	-0.002	-0.010	-0.4	0.725	
Region: South	-0.046	-0.220	-7.3	< 0.001	4.79
Region: West	-0.059	-0.176	-7.4	< 0.001	
log(Population Density)	-0.009	-0.137	-4.4	< 0.001	5.91
% Male	-0.879	-0.141	-7.0	< 0.001	2.43
% under 5 Years Old	-3.243	-0.305	-9.9	< 0.001	5.70
% over 65 Years Old	-0.222	-0.085	-2.5	0.012	6.77
% American Indian	-0.117	-0.065	-3.9	< 0.001	1.64
% Asian	0.505	0.075	4.1	< 0.001	2.00
% Native Born	0.209	0.084	4.0	< 0.001	2.64
% Never Married	0.124	0.062	2.1	0.039	5.34
Average Household Size	0.070	0.121	3.3	0.001	7.83
Avg Family - Household Size Delta	0.441	0.286	11.0	< 0.001	4.00
% Owner Occupied Housing	-0.001	-0.070	-2.7	0.006	3.89
% Enrolled in College	-71.090	-0.232	-10.5	< 0.001	2.90
% Higher Education Graduate	-0.181	-0.126	-5.0	< 0.001	3.86
Median Income	-0.002	-0.098	-4.4	< 0.001	2.90
% in Labor Force	-0.263	-0.169	-4.8	< 0.001	7.45
% of Labor Force in Armed Forces	1.017	0.135	8.9	< 0.001	1.38
% of Labor Force in Construction	-0.576	-0.125	-6.7	< 0.001	2.08
% of Labor Force in Retail	-0.573	-0.105	-6.8	< 0.001	1.44
% of Labor Force in Public Admin	-0.291	-0.076	-4.6	< 0.001	1.61
% of Labor in Finance/Real Estate	-0.428	-0.074	-3.9	< 0.001	2.12
% of Labor Force in Agriculture	0.097	0.064	2.5	0.013	3.93
% of Labor in Arts & Entertainment	-0.207	-0.058	-3.4	0.001	1.70
% of Individuals below Poverty Line	0.120	0.070	2.0	0.042	7.17

Figure 2. Residual Errors for Model Estimations

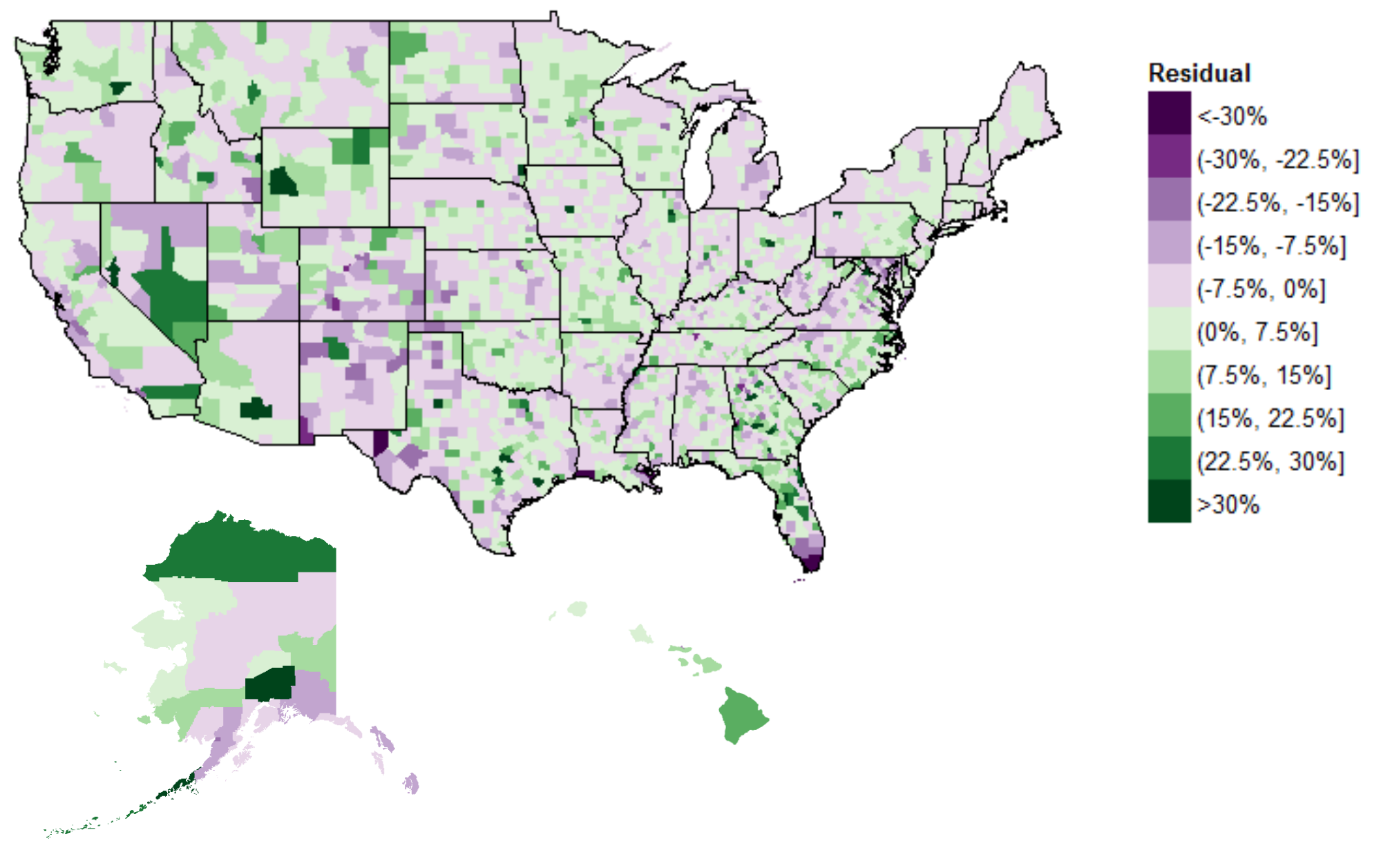


Table 3. Actual and Predicted Population Change for Select Observations

County Name	2000 Population	2010 Population	Population Change	95% Prediction Interval
St. Bernard Parish, LA	67,229	35,897	-46.6%	(-4.9%, 27.6%)
Cameron Parish, LA	9,991	6,839	-31.5%	(-9.2%, 20.1%)
Pinal County, AZ	179,727	375,770	109.1%	(1.8%, 40.2%)
Lake County, CO	7,812	7,310	-6.4%	(18.4%, 75.2%)
Sublette County, WY	5,920	10,247	73.1%	(-4.6%, 28.2%)
Bartow County, GA	76,019	100,157	31.8%	(6.0%, 47.9%)
Cherokee County, GA	141,903	214,346	51.1%	(23.0%, 83.5%)
Cobb County, GA	607,751	688,078	13.2%	(9.5%, 55.0%)
Fulton County, GA	816,006	920,581	12.8%	(-6.5%, 25.0%)
Forsyth County, GA	98,407	175,511	78.4%	(33.8%, 109.5%)
Gwinnett County, GA	588,448	805,321	36.9%	(14.9%, 66.3%)

RESULTS

As Table 1 shows, model performance was relatively similar for all forms considered. The first order OLS model was selected as the final model due to its relatively high $R^2_{predicted}$ value and a significant reduction in complexity compared to models with interactions. Regression was significant for the first order OLS model, *F(26, 2510)* = 133.8, $p < 2.2 \times 10^{-16}$.

All coefficient estimates for the final model can be found in Table 2. Bear in mind that a transformed response was used (the ratio of 2000 population to 2010 population); thus, a negative coefficient indicates a *positive* relationship between that variable and the actual 2000 to 2010 population change. Some of the most significant factors were as follows:

- Percent of population under 5 years old: more youngsters corresponded to more growth.
- The difference between average family size and average household size: this metric was created to resolve multi-collinearity issues between household size and family size, both of which were found to be significant. A higher size difference might indicate a larger single population in an area; larger differences corresponded to less growth.
- Percent of population enrolled in college: higher college enrollment in an area equated to greater population increase.
- Geographic region: the South and West had more population growth than the Midwest and Northeast.
- Percent of population over age 16 in the labor force: more people in the labor force meant more population growth.

Model performance by county can be assessed by the choropleth of residuals in Figure 2: purple corresponds to counties where model estimates were too high, and green corresponds to low estimates. Some of the most influential counties, as well as some local counties, are presented in Table 3 along with the corresponding prediction intervals. The most influential observation – St. Bernard Parish, Louisiana – was heavily impacted by Hurricane Katrina.

CONCLUSIONS

The intent of this research was to develop a model for population change in the United States between 2000 and 2010 using only decennial census information. A significant regression model was developed, and several key indicators, such as the proportion of the population under 5 years of age and the percent enrolled in college, were identified. This research suggests that including certain demographic, social, and economic variables in population models could be beneficial. Despite these successes, however, the utility of the model is limited as it only explains around 58% of the total variation in population change. Additionally, it is unclear how the model might perform at projecting forward in time. Could this model be applied to 2010 census data to predict 2020 population trends? This might be an area for further investigation.

R CODE

Code for Development and Validation of Final Model – OLS 1st order

```r
library(car); library(MASS)
OLS.1st.final <- lm(Ratio2 ~ Region + Male + Age.Under5 + Age.65.Plus +
       Race.AmerIndian + Race.Asian + Household.Size + House.Family.Delta +
       Own.Housing + HigherEduc.Grad + Never.Married + NativeBorn + LaborForce +
       Armed.Forces + Median.Income + Agri + Construction + Retail + Finance + Arts +
       PublicAdmin + Poverty.Indiv + log(Density) + College, data=train.data)
summary(OLS.1st.final) #Check significance of regressors
anova(OLS.1st.final, OLS.1st.r1) #Extra sum of squares to test model subset
vif(OLS.1st.final) #Check for multi-collinearity issues
plot(OLS.1st.final) #Residuals analysis – assess model validity
avPlots(OLS.1st.final) #Partial regression plots
summary(influence.measures(OLS.1st.final)) #Examine influential observations
#Assess model's predictive ability using test data set
predix.OLS.1st <- predict(OLS.1st.final, test.data) #Calculate predicted values
resid.OLS.1st <- test.data$Ratio2 - predix.OLS.1st #Calculate residuals
SST <- sum(test.data$Ratio2^2)- sum(test.data$Ratio2)^2/nrow(test.data) #Total SS
OLS.1st.R2pred <- 1-sum(resid.OLS.1st^2)/SST #R^2 = 1 – SSE/SST
```

Code for Creation of Choropleth for Actual Population Change in Continental U.S.

```r
library(ggplot2);library(RColorBrewer);library(maps);library(mapproj);library(rgdal)
#Define the groupings for population change and apply to main data frame
raw.cuts <- c(min(model.data$Change, na.rm=TRUE)-.0001, seq(-0.3, 0.375, by=0.075),
       max(model.data$Change, na.rm=TRUE)+0.0001)
model.data$raw.cuts <- cut(model.data$Change, raw.cuts)
buckets <- brewer.pal(11, "RdYlBu")[11:1] #Define color palette for graph
#Read in GIS/geographic data
counties <- readOGR(dsn=".",layer="gz_2010_us_050_00_5m")
#Create FIPS codes to match main data frame
counties$FIPS <- paste(counties$STATE, counties$COUNTY)
FIPS.convert <- counties@data
FIPS.convert$id <- rownames(FIPS.convert)
temp <- merge(model.data, FIPS.convert, by="FIPS", all=TRUE)
geo.temp <- fortify(counties) #Convert into data frame for plotting
geo.data <- merge(geo.temp, temp, by="id", all=TRUE) #Merge data frames
#Split out continental US
continental <- geo.data[!geo.data$STATE%in%c("02", "15", "72"),]
state.map <- map_data("state") #Data to map state boundaries
#Create continental US choropleth of actual population change
cont.raw <- ggplot(continental, aes(long, lat, group=group))+ coord_map() +
       geom_polygon(aes(fill=raw.cuts))+ scale_fill_manual(name="Population Change",
       values=buckets, labels=leg.labs) + geom_path(data = state.map, colour =
       "black", size = .5) + ggtitle("Figure 1. Actual Population Change in the
       United States from 2000 to 2010") + theme(panel.background=element_rect(fill=
       "transparent", color=NA), axis.ticks=element_blank(), axis.text.y=
       element_blank(), axis.text.x=element_blank(), axis.title.x=element_blank(),
       axis.title.y=element_blank())
```