Predicting Unemployment

Dirk Broadhead, Yuxin Chen, Jordan Okada, Audrey Woodwell

17 April 2020

1 Introduction

1.1 Unemployment

Widespread unemployment can severely harm the many aspects of the economy. All of the unemployed individuals lose their source of income and businesses are often unable to produce as many goods or services. Additionally, the purchasing power of the unemployed individuals is decreased, which often negatively impacts other businesses. By knowing the unemployment rate in a given community, civic leaders and policy makers can help determine the overall health and growth of the economy. A low unemployment rate is typically better because it implies that individuals who are searching for a job are more likely to find employment. The purpose of our study is to attempt to understand the relationship between unemployment and several of the other population indicators regularly collected by the United States government. Our study focuses specifically on the data collected from the state of Utah.

1.2 American Community Survey

The United States of America population census takes place every 10 years. However, taking a count of every individual in an entire year is a costly undertaking, so instead a sample survey of the population, known as the American Community Survey (ACS), is conducted every month by the Census Bureau. This survey is sent to a random sample of 3.5 million addresses across the United States. In this way, communities can get current estimates of various statistics for the population every year, rather than relying solely on the information from the last census. The ACS is also used by both local and national leaders to determine the need for further funding, programs, or other projects. For our analysis, we use data from the ACS in 2017 (*The Importance of the American Community Survey and the 2020 Census* (2020)).

1.3 Census Tracts

"Census tracts are relatively permanent small-area geographic divisions of a county or statistically equivalent entity defined for the tabulation and presentation of data" (*Census Tracts for the 2020*

Census-Final Criteria (2018)) We decided to use census tracts rather than cities or counties because a census tract allowed us to closely examine the population of a small geographic area within a county or city. Using data from smaller areas gives us more information to use in our analysis. For example, there are 588 census tract in Utah, which means we have more than 500 observations in our data set. However, there are only 29 counties in Utah, an extremely significant decrease in the number and diversity of the data set.

2 Data

Variable Name	Definition
Income	The average household income
Percent Men	Percent of Men in the total population. The complement is the percent of women.
Poverty Rate	Percent of population in poverty
Child Poverty Rate	Percent of children in poverty
Total Population	Number of people in each census tract
Mean Commute	Average work commute in a census tract
Proportion of Voting Age Citizens	Percent of population above voting age
Income Error	Median househould income error
Income Per Capita	The average income per person of a census tract

Figure 1: Variables Used in the Study

The data in the ACS contains information from across the United States. We first separated out the 588 census tracts that are located in Utah. Figure 1 shows the initial nine explanatory variables we began with. We used these variables to explain the unemployment rate by census tract in Utah.

There were three census tracts with no reported values other than the location of the tract (county and state). For the purpose of this analysis we removed these observations because they added no additional information.

Before beginning our analysis, we separated 20% of the observations (approximately 118 data points) from the main data set. 80% of the data was used for the creation of various inference and prediction models. The small portion of the observations that were removed was used later to test the validity of our potential models.

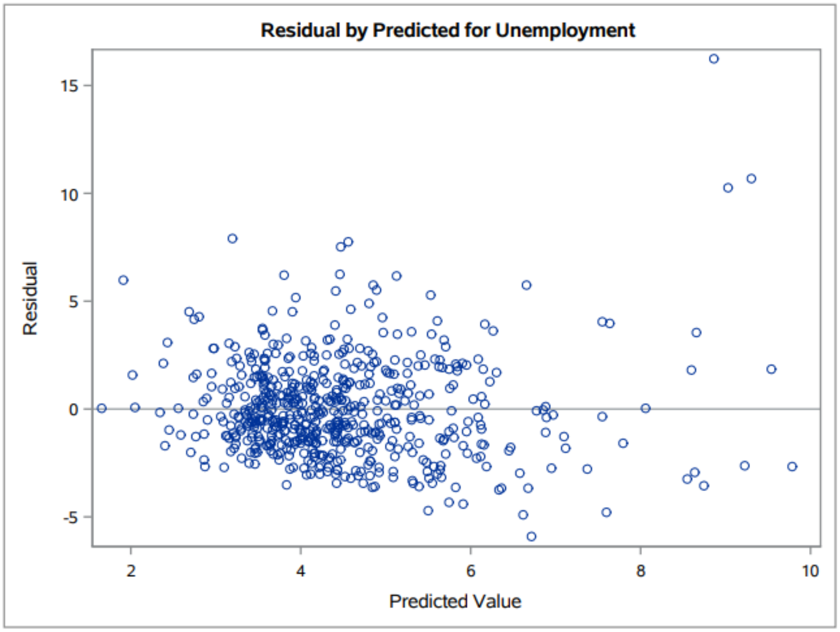

Figure 2: Residual Plot of unemployment

P-value for Brown-Forsythe test of constant variance in Residual vs. Predicted

Obs	t_BF	BF_pvalue
1	4.24265	.000026692

Figure 3: Brow-Forsythe Test of Constant Variance

The necessary conditions of a linear model appear to be violated. While the distribution of unemployment is approximately normal, there is evidence of slight non-constant variance. Figure 2 shows the distribution of the residuals against the predicted values, and Figure 3 is the p-value for the Brown Forsythe test of constant variance. Both of these outputs indicate that the residuals for unemployment have a heteroscedastic distribution. Taking a log transformation of unemployment sufficiently reduced the non-constant variance. The significant p-value for the Brown-Forsythe test of constant variance on the log transformed data is given below in Figure 4.

Obs	t_BF	BF_pvalue
1	1.33008	0.18419

Figure 4: Brown-Forsythe Test of Constant Variance

After performing a transformation on unemployment, we tested for multicollinearity in the independent variables. Total population is the sum of the number of men and the number of women and is an exact linear combination of those two variables. The variables for men, women, and total population had very high variance inflation factors (3195, 2895, and 12036 respectively). The condition index also reflected this multicollinearity. Additionally, the number of voting age citizens had a very high variance inflation factor and was highly collinear with men, women, and total population. Rather than deleting several variables, we transformed the way the variables were presented. We encoded the number of men as a percentage of total population. This automatically includes the data about the total number of women, because the percent of women in the total population is the complement of the percent of men. We also expressed the number of voting age citizens as a percent of the total population. Total population was left alone. This adequately reduced the variance inflation factors to be well below 10. The variance inflation factors for the other variables were also low, suggesting that none of the other variables were highly collinear.

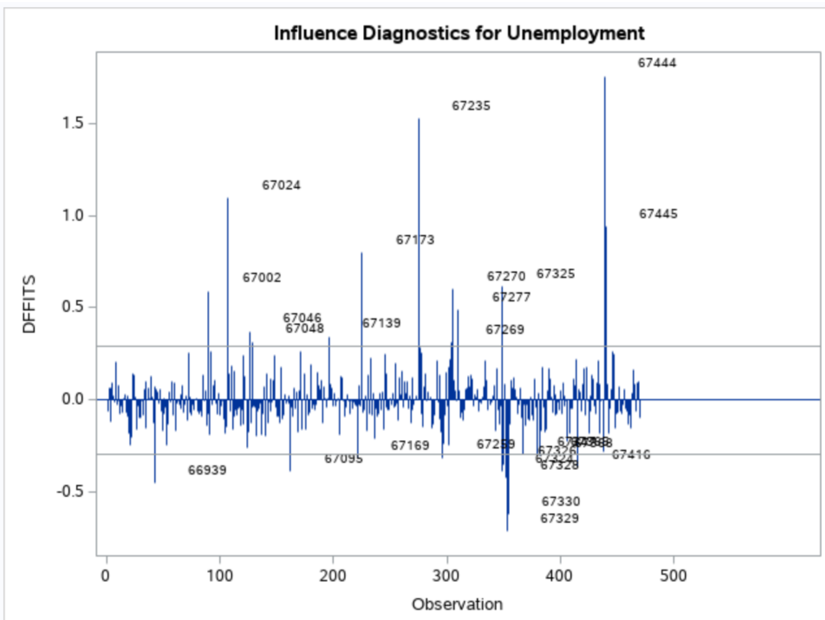

Figure 5: DFFITS Plot for Unemployment

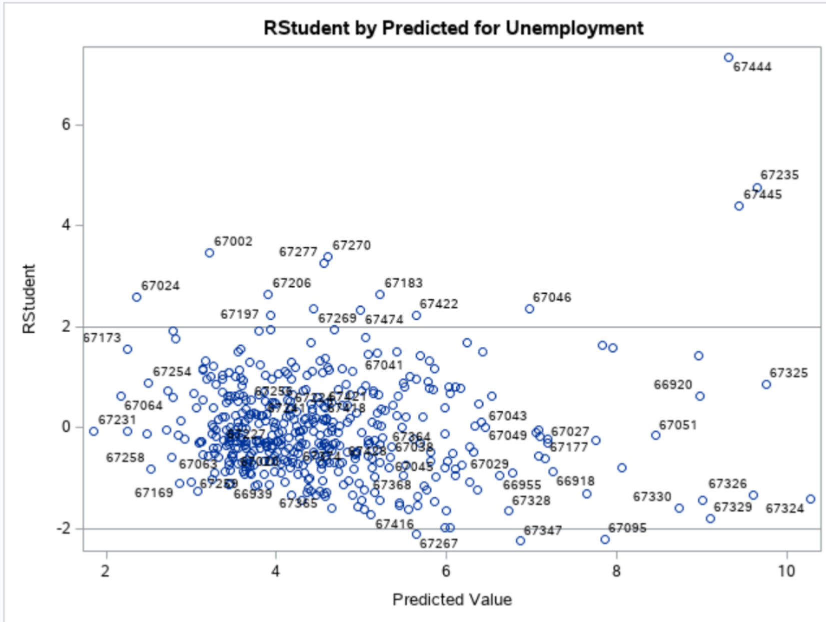

Figure 6: Studentized Residuals for Unemployment

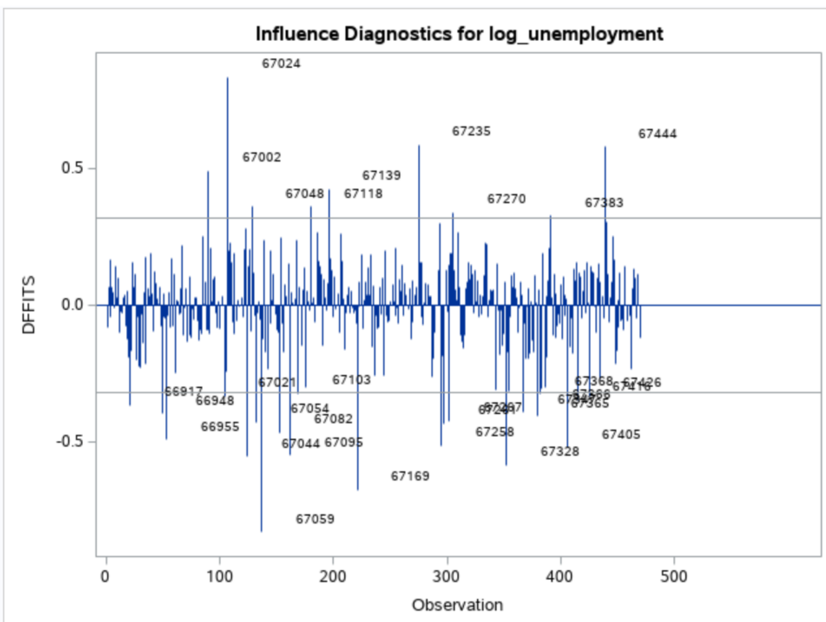

Figure 7: DFFITS for Log Transformed Unemployment

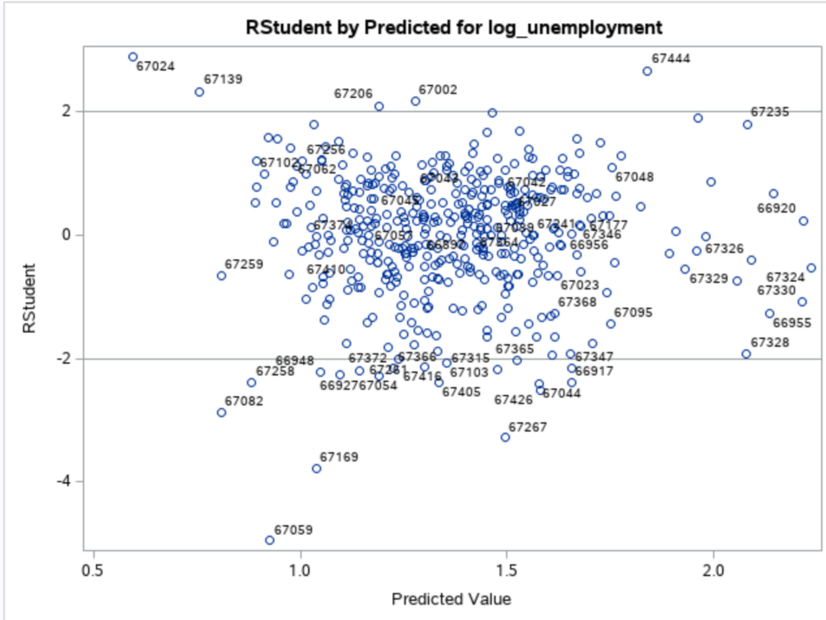

Figure 8: Studentized Residuals for Log Transformed Unemployment

There were several influential points that can be seen in the DFFITS plot in Figure 5. Additionally there were several outliers that can be seen in the plot of studentized residuals for unemployment in Figure 6. These influential points and outliers exist because most of our explanatory variables are right-skewed. Rather than removing so many observations, we transformed most of the explanatory variables. We used a log-transformation on most of the variables; with the exception of percent of men, percent of voting age citizens, and mean commute, which were almost exactly normally distributed to begin with. The distribution of the explanatory variables following the transformation was much closer to normal. And, while the transformation did not fully remove all of the influential points, it did sufficiently mitigate their effects. Figures 7 and 8 show an example of the reduced effects of both the influential points and the outliers following the transformation.

3 Model Testing

To determine the best model for prediction and inference using ordinary least squares regression, we used a variety of variable selection techniques. Both stepwise selection and backward selection suggest using a model with only two variables: poverty and income per capita. The best model based on the CP, adjusted R^2, AIC, and SBC in all possible regressions also included poverty and income per capita but suggested the addition of income error.

Our remedial measures did transform the data enough to satisfy the assumptions of a linear model. However, we also attempted to evaluate our data using a regression tree to see if a nonparametric method created better predictions. The tree had two branches and suggested income and the child poverty level were the two most important variables when making predictions.

4 Comparing Potential Models

In addition to the basic models suggested above we also tested to see if several interaction terms were significant. First, we tested to see if poverty and income per capita had any interaction. We tested to see if either poverty or income per capita had higher order interactions, specifically if they interact with themselves. The model with potential interaction terms that we tested was as follows:

$$\hat{Y} = \beta_0 + \beta_1(\text{income per capita}) + \beta_2(\text{poverty}) + \beta_{1,2}(\text{income per capita})(\text{poverty}) + \beta_3(\text{income per capita})^2 + \beta_4(\text{poverty})^2$$

The interactions were insignificant: (income per capita)(poverty) had a p-value of $p = 0.8243$, (income per capita)^2 had a p-value of $p = 0.7611$, and (poverty)^2 had a p-value of $p = 0.1151$. Because these interaction terms did not have a significant effect on unemployment, they will not be included in the model.

In order to determine which of the models performed the best, we calculated the ability of each model to make predictions on the separated 20% of the data. Using the test, set we calculated the mean square prediction rate (MSPR) for four linear regression models: one that contained no predictors, one that contained poverty and income per capita as predictors, a third that contained poverty, income per capita, and income error as predictors, and a fourth that contained all of the initial variables. We also calculated the MSPR for predictions made by the regression tree.

Model	MSPR
Tree	0.3657
2 Variables	0.3301
3 Variabes	0.3293
Full	0.3282
None	0.3672

Figure 9: Mean Squared Prediction Rate for Potential Models

As seen in Figure 9, the model with no predictor variables had the highest MSPR. The regression tree predicted only slightly better than the empty model, but there is not a significant difference. The full model and the models with two and three variables had very similar prediction error rates, that were approximately 0.329 (± 0.001). All three of these models decreased the prediction error by approximately 10%.

Because the models including two variables, three variables, and all nine initial variables all performed very similarly, we could justifiably choose any of of them for the final reported model. However, under the assumption that a simpler model is a better model, we will use the two variable model.

5 Final Model and Assumptions

The equation of our final model is given as:

$$\hat{y} = 5.13718 - 0.40302(\text{Income Per Capita}) + 0.14494(\text{Poverty})$$

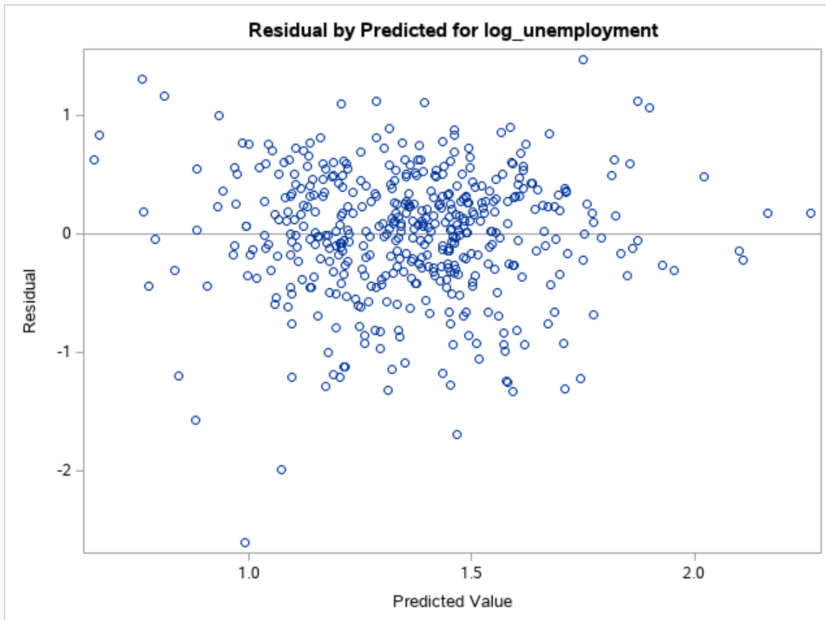

Figure 10: Residuals by Predicted Values

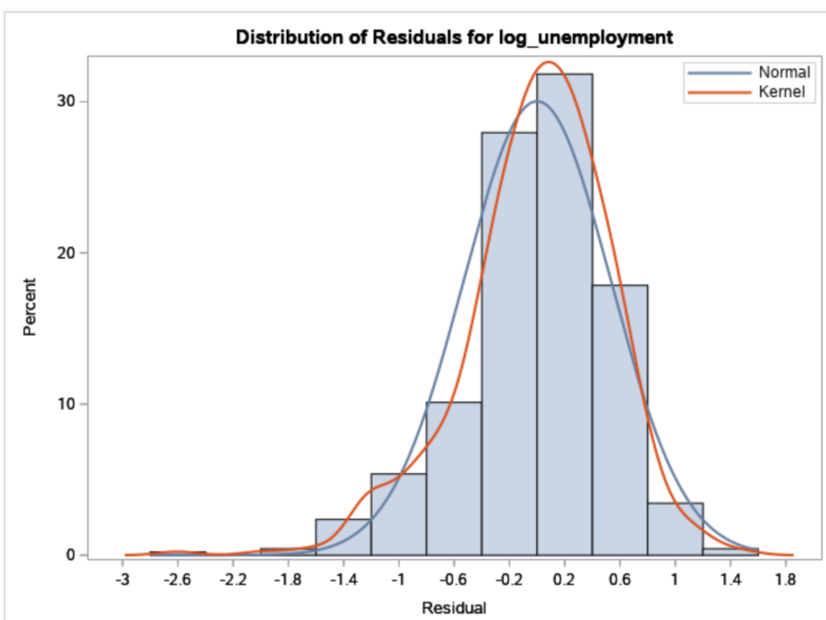

Figure 11: Distribution of Residuals

As seen in Figure 10 the distribution of residuals is homoscedastic, and Figure 11 shows that the distribution of the residuals is approximately normal. We can conclude that the assumptions for a linear model are satisfied.

Income per capita and unemployment have a negative relationship. After accounting for the effect of poverty for every unit increase in income per capita, the unemployment rate decreases by 0.40302. The influence of poverty on unemployment is positive. While holding the effect of income per capita constant for every unit increase in the poverty level, the unemployment rate increases by 0.14494.

6 Conclusion

Our predictions regarding the unemployment rate were not highly accurate. We are, however, able to infer several things about what influences the unemployment rate. We can conclude that the poverty rate is positively correlated with the unemployment rate; both rates increase together. Income per capita on the other hand is negatively related to the unemployment rate, meaning that as one of the two increases, the other decreases. Intuitively, the aforementioned relationships make sense. An increase in the poverty level means people are making less money and are more likely to be unemployed. However, as the average amount of money people in a census tract make increases it is likely that more individuals are actually employed.

We only studied the census tracts that are located in Utah. However, the same information is collected annually across all of the United States. Further study might attempt to generalize our techniques, to predict the unemployment of the United States as a whole, rather than just an individual state.

It would also be beneficial to see if our prediction of unemployment could be improved by adding different, potentially more applicable, predictor variables. Some of these variables might be a measure of the quality of K-12 schools, the average educational attainment in the labor force, or the average age of the labor force. Additionally future studies might use a different, more applicable, study to collect data. The ACS is beneficial because it contains a massive amount of data, it happens yearly, and the information is free to the public. However, it could be beneficial in future studies to collect information with the purpose of studying unemployment.

References

Census tracts for the 2020 census-final criteria. (2018, October 30). U.S. Census Bureau. Retrieved 2020-03-30, from https://www.federalregister.gov/documents/2018/11/13/2018-24567/census-tracts-for-the-2020-census-final-criteria

The importance of the american community survey and the 2020 census. (2020, February). Retrieved 2020-03-28, from https://www.census.gov/programs-surveys/acs/about/acs-and-census.html

7 Appendix SAS Code

```
    FILENAME REFFILE '/home/u42026342/STAT5100/Final Paper/utahCensus17.csv';

PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.census;
GETNAMES=YES;
RUN;

/*Initial Check of Data */
proc sgscatter data=census;
matrix unemployment income men women TotalPop VotingAgeCitizen Poverty ChildPoverty
MeanCommute IncomeErr IncomeperCap / markerattrs = (symbol=circlefilled size=2pt);
title1 'Scatter Plot Matrix of Variables';
run;

proc univariate data=census;
var unemployment income men women TotalPop VotingAgeCitizen Poverty ChildPoverty
MeanCommute IncomeErr IncomeperCap;
hist unemployment income men women TotalPop VotingAgeCitizen Poverty ChildPoverty
MeanCommute IncomeErr IncomeperCap;
run;

/*Fixing Multi Collinearity, Influential Point and Outliers, and Training set */
data census1; set census1;
percentment= men/TotalPop;
percentVoters=VotingAgeCitizen/TotalPop;

proc surveyselect data=census1 seed=1337 out=census rate=0.2 outall;
/* Withold 20% for validation */
run;

data train;
set census;
if Selected=0;
run;
data test;
set census;
if Selected=1;
run;
data train; set train;
data train; set train;
log_unemployment=log(unemployment);
log_men=log(men);
```

```
log_women=log(women);
log_childpov=log(ChildPoverty);
log_population=log(TotalPop);
log_poverty=log(Poverty);
log_income=log(income);
log_VotingAgeCitizen=log(VotingAgeCitizen);
log_IncomeErr =log(IncomeErr);
log_IncomeperCap=log(IncomeperCap);
run;

proc reg data=train plots(label unpack)=(cooksd Rstudentbyleverage dffits dfbetas);
id var1;
model unemployment = income men women TotalPop VotingAgeCitizen Poverty ChildPoverty
MeanCommute IncomeErr IncomeperCap;
output out=out2 r=resid p=pred;
title1 'Initial Regression Model Predicting Unemployment';
title2 'Also Influential Points and Outliers';
run;
%resid_num_diag(dataset=out2, datavar=resid, label='Residual', predvar=pred,
predlabel='Predicted');

proc univariate data=train;
variable log_unemployment log_income percentment TotalPop percentVoters log_poverty log_
MeanCommute log_IncomeErr log_IncomeperCap;
histogram log_unemployment log_income log_men log_women log_population log_VotingAgeCiti
MeanCommute log_IncomeErr log_IncomeperCap;
run;

proc reg data=train plots(label unpack)=(cooksd Rstudentbyleverage dffits dfbetas);
id VAR1;
model log_unemployment = income men women TotalPop VotingAgeCitizen Poverty ChildPoverty
MeanCommute IncomeErr IncomeperCap; vif collin;
output out=out4 r=resid p=pred;
title1 'Initial Regression Model Predicting lpg_Unemployment';
run;
%resid_num_diag(dataset=out4, datavar=resid, label='Residual', predvar=pred,
predlabel='Predicted');

proc reg data=train;
id VAR1;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap / vif collin;
output out=out3 r=resid p=pred;
title1 'Initial Regression Model Predicting Unemployment';
run;
```

```
%resid_num_diag(dataset=out3, datavar=resid, label='Residual', predvar=pred,
predlabel='Predicted');

/* Variable Selection */
proc reg data=train;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap / selection=cp adjrsq aic sbc;
title1 'All possible regression - Variable Selection';
run;

proc reg data=train;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap / selection=stepwise slentry=0.1 slstay=0.1;
title1 'Stepwise Selection - Variable Selection';
run;

proc reg data=train;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap / selection=backward slstay=0.1;
title1 'Backward Selection - Variable Selection';

proc reg data= train plots(unpack)=diagnostics;
model log_unemployment = log_IncomeperCap log_poverty;
output out=out1 r=resid p=pred;
title1 'Potential Final Model';
run;
%resid_num_diag(dataset=out1, datavar=resid, label='Residual', predvar=pred,
predlabel='Predicted');

/* TESTING INTERACTIONS */
data train; set train;
incomexpoverty= log_incomeperCap*log_poverty;
incomesquare = (log_incomeperCap)**2;
povertysquare= (log_poverty)**2;

proc reg data=train;
model log_unemployment = log_IncomeperCap log_poverty incomexpoverty incomesquare povert
run;

/* Regression Tree */
proc hpsplit data=train seed=15531;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap;
code file='/home/u42026342/STAT5100/Final Paper/regressiontree.sas/'; /* This saves the
run;
```

```
/* Validation */
data test; set test;
log_unemployment=log(unemployment);
log_men=log(men);
log_women=log(women);
log_childpov=log(ChildPoverty);
log_population=log(TotalPop);
log_poverty=log(Poverty);
log_income=log(income);
log_VotingAgeCitizen=log(VotingAgeCitizen);
log_IncomeErr =log(IncomeErr);
log_IncomeperCap=log(IncomeperCap);
run;

proc reg data= train noprint;
model log_unemployment = log_IncomeperCap log_poverty;
store VarSelectModel;
run;

proc reg data= train noprint;
model log_unemployment = log_IncomeperCap log_poverty log_IncomeErr;
store VarSelectModel3;
run;

proc reg data=train noprint;
model log_unemployment = log_income percentment percentVoters log_population log_poverty
MeanCommute log_IncomeErr log_IncomeperCap;
store FullModel;
run;

/* Regression Tree */
proc hpsplit data=train seed=15531 noprint;
model log_unemployment = log_income percentment log_population percentvoters log_poverty
MeanCommute log_IncomeErr log_IncomeperCap;
code file='/home/u42026342/STAT5100/Final Paper/regressiontree.sas/'; /* This saves the
run;

proc reg data=train noprint;
model log_unemployment = ;
store emptyModel;
run;

proc plm restore=VarSelectModel;
score data=test out=newTest predicted;
```

```
run;

proc plm restore=VarSelectModel3;
score data=test out=newTest3var predicted;
run;

proc plm restore=FullModel;
score data=test out=newTest2 predicted;
run;

proc plm restore=emptyModel;
score data=test out=newTest3 predicted;
run;

data newTest; set newTest;
ASE = (log_unemployment - predicted)**2;
run;

data newTest3var; set newTest3var;
ASE = (log_unemployment - predicted)**2;
run;

data newTest2; set newTest2;
ASE = (log_unemployment - predicted)**2;
run;

data newTest3; set newTest3;
ASE = (log_unemployment - predicted)**2;
run;

data scored;
set test;
%include '/home/u42026342/STAT5100/Final Paper/regressiontree.sas/';
run;

data testTree;
set scored;
ASE = (log_unemployment - P_log_unemployment)**2;
run;
proc means data = testTree;
var ASE;
title1 'From Regression Tree';
run;

proc means data= newTest;
```

```
var ASE;
title1 'With 2 Variables';
title2 'From Variable Selection';
run;

proc means data= newTest3var;
var ASE;
title1 'With 3 Variables';
title2 'From Variable Selection';
run;

proc means data= newTest2;
var ASE;
title1 'Model with All variables';
run;

proc means data= newTest3;
var ASE;
title1 'Model with no Variables';
run;
```